
User Participation and Community Formation in Peer Production Systems

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera Italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by
Giovanni Luca Ciampaglia

under the supervision of
Prof. Luca Gambardella, Dr. Paolo Giordano,
and Dr. Alberto Vancheri

December 2011
Rev. 08972ee

Dissertation Committee

Prof. Dr. Michele Parrinello	ETH Zürich and University of Lugano, Switzerland
Prof. Dr. Fabio Crestani	University of Lugano, Switzerland
Prof. Dr. Kristina Lerman	University of Southern California, USA
Prof. Dr. Santo Fortunato	Aalto University, Finland

Dissertation accepted on 15 December 2011

Research Advisor	Co-Advisor	Co-Advisor
Prof. Luca Gambardella	Dr. Paolo Giordano	Dr. Alberto Vancheri

PhD Program Director
Prof. Dr. Antonio Carzaniga

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Giovanni Luca Ciampaglia
Lugano, 15 December 2011

*Par-delà ce village, d'autres
villages, par-delà cette abbaye,
d'autres abbayes, par-delà cette
forteresse, d'autres forteresses. Et
dans chacun de ces châteaux
d'idées, de ces mesures d'opinions
superposés aux mesures de bois et
aux châteaux de pierre, la vie
emmure les fous et ouvre un
pertuis aux sages.*

Marguerite Yourcenar,
"L'Œuvre au Noir"

Abstract

This thesis explores the phenomenon of user participation in peer production communities. In large-scale collaboration communities such as wikis, open source software development teams, and file-sharing groups, members are, in general, not remunerated for contributing to the digital common, so the incentives to participate in these projects are usually construed in terms of extrinsic benefits such as building reputations, reciprocity, gratification, and providing a sense of membership to a socially cohesive group. In fact, like any social group, a community of commons-based peer production is endowed with its own norms, beliefs, and cultural features. In these systems such features, however, are themselves the result of a bottom-up process of cultural formation and opinion aggregation, mostly related to various aspects of the social production of digital content.

Unfortunately it is still poorly understood how this process happens – considering that most interactions between members are mediated by the digital artifacts (encyclopedic articles, source code, etc.) that the project is producing. To this end, we studied the process of community formation in a large peer production community by means of statistical analysis and agent-based modeling and simulation.

In the first part of this thesis we analyzed the activity of registered editors from the communities of five of the largest versions of Wikipedia, the free online encyclopedia. We found that the distribution of the user activity lifespan is distributed according to a mixture of log-normal distributions.

In order to understand these empirical patterns we developed, in the second part of the thesis, an agent-based model of a peer production community. Social influence is modeled in terms of dyadic user-page interactions, under the form of the bounded confidence rule. Thus a measurable aspect of user participation – activity lifespan – is linked to microscopic features of the dynamics of social influence. In order to study the behavior of our model, we perform a factor screening via global sensitivity analysis. We then calibrate our agent-based model using the empirical data from Wikipedia. To this end, an indirect

inference technique is devised and tested.

In light of the results of our model, in the third – and final – part of the thesis we analyze a recent dataset from the English Wikipedia and perform a longitudinal study of the life cycle of user activity. Using a non-parametric approach we study how the daily rate of editing changes during the lifespan of editors, and find a strongly inhomogeneous temporal life cycle. This approach enables us to look at the temporal evolution of editing activity for the whole community of Wikipedia editors.

These results suggest that user participation in peer production systems can be construed primarily as a process of mediated social influence and that other factors, most surprisingly intrinsic motivation to contribution, are less important in determining the overall activity lifespan of an individual. In conclusion this thesis shows how social simulation can be supplemented with large-scale data analysis in order to develop an empirically grounded approach to the computational study of collective social phenomena and, in particular, social computing platforms.

Acknowledgements

This thesis would not have been possible without the support of a number of individuals, who helped me, in one way or another, during the course of my doctoral studies. I would like to thank my supervisors Paolo Giordano and Alberto Vancheri for their support during all my studies, for providing an intellectually stimulating environment, and for giving me almost total freedom in the pursuit of my scientific interests. I would like to thank Luca Maria Gambardella for accepting to be my supervisor at the Faculty of Informatics, thus allowing me to enroll in the doctoral program. I am grateful to my doctoral committee – Professors Michele Parrinello, Fabio Crestani, Kristina Lerman, and Santo Fortunato – for lending their time serving in it and for their valuable comments on the dissertation proposal. I am indebted to Amirhossein Malekpour, Prof. Fernando Pedone, and Prof. Antonio Carzaniga for allowing me to run simulations on their servers. Part of the research of this dissertation was performed during a summer fellowship at the Wikimedia Foundation, so I would like to thank Dario Taraborelli, Diederik Van Liere and Zack Exley on behalf of all the people there. I would also like to thank Prof. Dirk Helbing for hosting me at the Chair of Sociology at ETH Zürich during the winter of 2009 and all the members of his team.

Finally, I would like to thank (in no particular order): Alessandra Gorla, Cyrus Hall, Jeff Rose, Shane Legg, Nicolas Schiper, Mircea Lungu, Daan Wierstra, Tom Schaul, Matteo Gagliolo, Mark Carman, Julian Togelius, and Rachid Rebiha, for the stimulating discussions and, in general, for sharing four years of life in Lugano with me. The above list, of course, is not meant to be exhaustive: life is too large to be compressed in written form anyway.

Contents

Contents	ix
1 Introduction	1
1.1 What is peer production and why should we care about it?	1
1.2 Motivation	3
1.2.1 Wikipedia	3
1.2.2 Software debugging is parallelizable	5
1.2.3 How do peer production communities form?	7
1.3 Research questions	8
1.3.1 The general research statement	8
1.3.2 Operational research question	10
1.4 Main contributions of this dissertation	13
2 Literature survey	15
2.1 Structure of this chapter	16
2.2 Social informatics and social computing	17
2.2.1 Social Informatics	17
2.2.2 Social computing	21
2.3 The dynamics of collective social phenomena	24
2.3.1 Physics and society	24
2.3.2 Complex Networks	25
2.3.3 Opinion dynamics	27
2.3.4 Human dynamics	32
2.4 The socio-economics of online communities	33
2.4.1 Incentives to contribution	33
2.4.2 Social psychology of user participation	37
2.5 The research on Wikipedia	40
2.5.1 Wikipedia as a case study on commons-based peer produc- tion	40

2.5.2	Governance and work organization	40
2.5.3	Quality and the epistemic problems of peer production . .	42
3	Empirical analysis of user activity lifespan	43
3.1	Introduction	43
3.1.1	Empirical findings	44
3.2	The dataset	45
3.2.1	Data collection	45
3.2.2	Statistical considerations	47
3.2.3	Robot users	48
3.3	The Model	48
3.4	Results	51
3.4.1	Multi-modality of activity lifespan distribution	51
3.4.2	Inactivity periods	56
3.4.3	Temporal evolution	60
3.5	Discussion	61
4	Formation of peer production communities	65
4.1	Group formation in online communities	65
4.2	Model description	67
4.2.1	Dynamics of cultural traits	67
4.2.2	Editing model	68
4.2.3	Population dynamics	70
4.3	Model implementation	71
4.4	Simulation of cultural dynamics	72
4.5	Discussion	73
5	Factor screening and sensitivity analysis	77
5.1	Introduction	77
5.2	Methods	78
5.2.1	Local sensitivity analysis	78
5.2.2	Three approaches to global sensitivity analysis	79
5.2.3	Experimental design	84
5.3	Results	87
5.3.1	Simulation scenario	87
5.3.2	Factor screening via global sensitivity analysis	90
5.4	Discussion	95

6	Simulation-based model calibration	99
6.1	Introduction	99
6.2	Background: Indirect inference	100
6.3	Indirect inference of unknown density	102
6.4	Simulation design	104
6.4.1	Time scales of user lifespan	107
6.4.2	Activity rates	112
6.5	Simulation and Diagnostics	112
6.5.1	Approximation of auxiliary parameters via Gaussian Processes	113
6.5.2	Sensitivity analysis of auxiliary parameters	116
6.5.3	Cross-validation	117
6.6	Results	131
7	The life cycle of user activity	135
7.1	Introduction	135
7.2	Methods	136
7.2.1	Data Collection	137
7.3	Results and discussions	139
7.3.1	Editor productivity	139
7.3.2	User activity trends	142
7.3.3	Time scale of activity decay	144
7.3.4	Community dynamics	146
8	Conclusions	155
8.1	The distribution of the lifespan of user activity	155
8.1.1	Temporal trends of user activity	157
8.2	An agent-based approach to modeling peer production	157
8.2.1	Beyond bounded confidence and other developments	158
8.3	Future works and generalizations of the present work	159
	Bibliography	163
	Index	191

Chapter 1

Introduction

The problem with Wikipedia is that it only works in practice. In theory, it can never work.

The Zeroth Law of Wikipedia

1.1 What is peer production and why should we care about it?

This thesis deals with a specific question about the development of sociality in an online context: how people, Internet users, who altruistically build certain intangible artifacts, in doing so provide themselves with rules, norms, and customs – in practice, how they come to form a community of peer production. This problem has puzzled scholars since the phenomenon of mass collaboration blossomed on the Internet at the beginning of the last decade.

The term “commons-based peer production” (from now on: peer production) refers to the phenomenon of decentralized, loosely organized, groups of people rallying around projects that involves the production of information goods; this pool of shared information is said to form a digital common.¹ Such teams usually have a horizontal structure, weak leadership, and almost non-existent management: hence the denomination of “peers” for their participants. Groups are of varying size, ranging from a handful to the hundreds of thousands.

¹The term “common” refers to the shared grazing grounds that villages in England, some parts of northern Europe and, later, the USA, used to have. The custom of retaining free access to these areas dates back to the Middle Age, and managed to survive until late 19th century.

Why would peer production be an interesting subject of research? The first reason is that the phenomenon of mass collaboration online is still poorly understood. This, though, seems to clash with the reality of things: people are, nowadays, accustomed to collaborating on the Internet. They are used to participating in a disparate array of discussion groups, viewing and sharing photos and video from content-sharing sites, and querying question & answer communities or online encyclopedias – just to give a few examples. So, in a sense, people seem just as naturally drawn towards peer production as with any other pro-social behavior characteristic of human cooperation.

But despite man being – as the saying goes – a social animal, the online setting poses incredible challenges to the development of cooperative behavior: contributors are often anonymous; there is no a priori way to assess the accuracy of information; communications are asynchronous. In order to build a high-quality intellectual product, in fact, common sense would dictate that one should rather raise the barrier of contribution.

Thus, for a long time, it was thought that successful mass collaboration was very hard, if not downright impossible, to achieve. Instead, by being inclusive rather than exclusive, by lowering the barriers of contribution, and by abating transaction costs, peer production systems proved that it is possible to create huge information repositories in a very short time.

The second reason is that peer production is redefining the way knowledge is produced in modern societies. A central tenet of knowledge production is expertise. Now we know that it is possible to build an encyclopedia without the centralized control of a group of experts.

Of course this is not to say that, after Wikipedia, expertise can be considered dead: Sanger [2009] notes that, from an epistemological standpoint, it is a paradox to say that (a perfect version of) Wikipedia could replace experts, because one would need experts to assess its quality in the first place.

A clarification is needed at this point. With the present research we are not interested in construing the motivations people have for cooperating in such enterprises. Nor are we interested in assessing the value of information produced by peer productions efforts.

What is still not clear, and which this dissertation wishes to address, is how such large collaborative enterprises come about and thrive, how people coordinate their efforts and stick around for longer periods than any conceivable initial dose of enthusiasm could account for, and how they actually come to produce a coherent intellectual product and not just a disparate collection of incompatible contributions – be it an encyclopedia, an operative system kernel, or a virtual word experience.

In the rest of this chapter I will describe peer production through two famous examples: Wikipedia, and Free Open Source Software. I will highlight the strength and weaknesses of peer production, and argue that understanding how peer production works is interesting both for the social scientist and for the computer (and information) scientist. In particular, we shall focus on the two related problems of user participation and community formation, that is, roughly speaking, what drives people to collaborate for free on a project, and how their joining forces can provide an incentive for further participation in the collective endeavor. I will, at this point, formulate the research questions of this thesis.

1.2 Motivation

1.2.1 Wikipedia

According to a survey by the Pew institute, in 2011 42% of all adult Americans turned to Wikipedia for information online (cf. Zickuhr and Rainie [2011]). Wikipedia usually ranks 6th or 7th among the top most popular Internet websites, according to Alexa.com [2011]. According to statistics of the Wikimedia Foundation (the non-profit organization that manages its infrastructure), its reader base has been reported to be 381 million unique visitors worldwide, as of April 2011 (cf. Zachte [2011]).

These numbers look impressive, but how has Wikipedia managed to achieve such popularity? The project was founded by Jimmy Wales as a spin-off of Nupedia, an earlier project that wanted to build an open online encyclopedia with the contributions of volunteer experts, using a traditional editorial process.

The Wiki technology provides two main benefits: it reduces the transaction costs of contribution, since users just need to click on a button in order to edit any page of the website; and it is incremental, in the sense that any version of the page is accessible for later review. Therefore it is easy to restore the contents of a page to a previous, safe state, if somebody inserts an error or removes legitimate content. Besides these two features, the Wiki technology does not provide any real feature to check the accuracy of what is introduced ²

According to Reagle [2007b], Wikipedia follows the same tradition of several utopian projects, the earliest of which date back as far as the early 20th century,

²To be precise, there is an extension to the Mediawiki software – the Wiki engine that powers Wikipedia – called edit filter. This extension allows administrators to specify some generic, hard-logic rules for dealing with abusive contributions. An example of a rule is: “*users with fewer than 500 edits are blocked from moving pages to titles which match this regular expression: /poop/*”, cf. <http://bit.ly/q8TM1k>.

like the World Encyclopedia – a repository of all human knowledge envisioned by writer H.G. Wells [1938]. However, the reason why Wikipedia has succeeded is, again according to Reagle [2007b], because of its collaborative culture.

A set of policies regulates nearly all aspects of the collaborative process of Wikipedia. The normative body itself is also the result of a bottom-up process of distillation of informal practices and customs. At the core, the most important policies state that:

- a. Wikipedia works by consensus; interactions should always be based upon civility and good manners, and editors should assume that their peers always act in good faith.
- b. Entries should be written following the so-called Neutral point of view (NPOV). This means that all existing major points of view on a given topic should be portrayed in the entry dedicated to that topic. When writing on a specific subject, rather than advocating a specific stance on it, an editor should try to describe that stance.
- c. Information inserted in an entry should always rely on authoritative, external sources on the subject.

We can see already, from this group of core policies, that the philosophy of Wikipedia is highly inclusive. The NPOV rule is key here. Writing in a neutral style means that all points of view on a given topic may be described, as long as they are perceived to be relevant by the public. Validity is just a subordinate criterion. Hence an article about evolution can contain information about the creationist interpretation of the theory of Darwin, as long as this information is limited to a description of the point of view itself, and does not argue its validity. The distinction is often subtle: describing a belief is often just a way to argue in favor of its truth after all.

The other tenet of Wikipedia is that authority shifts from editors to sources. While this greatly simplifies disputes among editors, it is also the main limitation of Wikipedia, since editors who are not experts on a topic often confuse the popularity of a source for its accuracy. Indeed many academics who have participated in Wikipedia lament having lost interest in it, usually after having battled over some detail with one or more non-expert fellow editors, who backed their claims using inaccurate sources taken from Google or other search engines – and who had much more free time than them.

It is no surprise, then, that the main concerns expressed against Wikipedia are about the accuracy of the information inserted. It is perhaps after the study

commissioned by the reputable journal *Nature* (cf. Giles [2005]) that Wikipedia earned a good deal of fame within academic circles. The study performed a blind peer-review comparison between articles taken from Wikipedia and from the website of *Encyclopædia Britannica*, on a selection of topics ranging from the natural sciences to the humanities. The results *Nature* reported were strikingly encouraging for Wikipedia: out of 42 reviews, on average *Britannica* scored 3 errors per entry, while Wikipedia 4. The methodology and the conclusions of the study, however, have been disputed by *Britannica* (cf. *Encyclopædia Britannica* [2006]), calling for a retraction from *Nature*. Ironically, the value of this story lies perhaps in the fact that evaluating the quality of reference works is an intrinsically hard task.

1.2.2 Software debugging is parallelizable

Peer production is a very recent phenomenon, but how did it come to be so popular? The first noteworthy experience that showed that producing high-quality products following such a model was possible came from the world of information technology: it was the Free Open Source Software (FOSS) movement.

Even though different elements of the open source ideology have been implemented in different ways, the general idea behind it is that software – the product that, in a sense, is being manufactured – should be freely accessible, modifiable, and redistributable by anybody.³ This is true both for source and binary code. Under this type of distribution license, in particular, any copyrights on the software are essentially waived.⁴

As we have seen, from a purely philosophical standpoint this is certainly appealing, but why is this openness a good thing from a practical point of view? The main advantage of open source software is its resiliency to bugs. And the reason for this lies in the advantage of harnessing peer production. Indeed, every piece of software, free or proprietary, should be understood as “living” in a social space in which developers interact with their user base.

For firms, the user base is composed of its customers, who are only supposed to use the software and, ideally, provide some feedback about new features,

³In addition to this, software licensed under the GNU General Public License (GPL), adds a “viral” clause that forbids from redistributing under stricter terms than those of the GPL. This clause does not hold generally for Open Source software, hence the distinction of terminology between Free Software (i.e. GPL’ed) and Open Source. A widely used Open Source license is the BSD license.

⁴Even though it should be noted that the copyright is still asserted by the author of a piece of software; users are then expressly granted the specific rights to copy, modify, and re-distribute. This is the technical mechanism, at the legal level, that the GPL for example uses.

bugs, and security issues. The phases of design, testing, and implementation, are entirely performed within the firm.

However, here come the problems. The firm has to figure out what context the software is used in, if it wants to test it thoroughly: what kind of operative system the software runs in, on what kind of hardware, etc. It is empirically known that these parameters, the combination of all possible use cases and platform requirements, scale with the size of the user base much faster than the resource of any firm can keep up with. This means that, in order to produce a reliable piece of software, a firm must, before releasing its product to the public, perform expensive and time-consuming cycles of testing and debugging.

This is not the case in the open source world. Software is released early and often, without any guarantee of being free from bugs. Testing and debugging are performed by distributing the scrutiny of the code across large groups of volunteers – the user base. It is normal for users of open source software to report problems on specific mailing lists and other online communication venues, such as Internet Relay Chat (IRC) chat rooms, or bug-tracking databases. These are, essentially, meeting points between end-users and developers. People go there to seek assistance with the software; at the same time developers use them to receive feedback – not just bugs, but also requests of new features – and to announce new versions of the software.

The existence of these communication platforms – Raymond [1999] suggestively compares them to noisy bazaars, and software firms to austere cathedrals – might raise some concerns about how developers can keep up with so many requests. The reason is simple: because source code is freely available, other programmers can propose modifications (“patches”, in the technical jargon) to fix problems they discover. Or they can take entire chunks of code and rewrite them completely – for example to enhance their performance – or introduce new features. This social distribution of work greatly reduces the load on the core maintainers of the software.

In short, an open source community is a thriving social space where end-users are an integral part of the development cycle. Compared to proprietary software, where code undergoes long and structured testing cycles, this social space offers a solution to the problem of how to deal with software complexity. Of course this is not to say that the code developed by volunteer programmers is bug-free, but simply that open source offers a more reactive development model compared to the proprietary one. When the source code is available for any user to inspect, suddenly it undergoes the scrutiny of several hundreds, if not thousands of individuals. Linus Torvalds, the creator of Linux, popularized this concept by saying, “*given enough eyeballs, all bugs are shallow*”, and Raymond

[1999] reformulated it as: “*debugging is parallelizable*”.

1.2.3 How do peer production communities form?

This chapter opens with a funny but thought-provocative quote. So far, the general opinion holds that Wikipedia works in practice, but it cannot work in theory. Does it mean that projects like Wikipedia are just lucky accidents of history? Or can we try to understand the “theory” behind a peer production community? Can we rationalize what ingredients and mechanisms are responsible for the development of a healthy community? And what may lead, instead, to its demise?

In fact, the Internet is full of dead, inactive communities. And they are not all software projects or wikis. Social networks, news filtering websites; discussion groups; USENET newsgroups, IRC chats; massive multi-player online role-playing games (MMORPG, the successors of early multi-user dungeons, or MUDs): all these have similar incentive structures. So far, any explanation of why all these communities become extinct has almost invariably revolved around exogenous factors, like the decline in popularity, the rise of external competition, etc. Of course these are sensible considerations, but we cannot help noting that very few commentators take into account endogenous factors to explain this phenomenon.

A different rationale would be that, whenever people contribute to an open project because they feel the urge to belong to a healthy community, we should, in the first place, understand what makes for a healthy community. We argue that the reason why this has not been the case, at least at a quantitative level, lies in the absence of a complete microscopic framework that takes into account incentives of the people and the nature of the activity they are engaged in.

The study of motivation for contributing to FOSS and to Wikipedia gives us some insight into how peer production works, but it also reveals that most of these explanations rely on factors, such as the sense of belonging to a community, which are still largely unexplained. Benkler [2002, p. 424] exemplifies this state of affairs:

What makes contributors to peer production enterprises tick? Why do they contribute? [...] It would seek to understand the motivation and patterns of clustering around projects in the absence of property rights and contracts and the emergence of the effective networks of peers necessary to make a *particular* project succeed. These are questions that present rich grounds both for theoretical and empirical study.

Ideally, a particular incentive structure would be the reason, as well as the outcome, of a process of “clustering” that happens at the level of the group of contributors. But this is a classic idea from the physics of complex systems, and in particular statistical physics: the dynamics at the microscopic level can lead to the emergence of patterns at the macroscopic level. Of course here we are not dealing with a gas, but with people. What kind of rules govern the interaction between people? Social psychology can provide us with the right framework for describing this microscopic level of interactions between peers. This brings us to the subject of this research.

1.3 Research questions

The present dissertation is an attempt to give a more concrete foundation to the science of large-scale peer production systems. In particular, we are interested in a quantitative foundation, and in comparing our predictions with empirical data from existing online communities of peer production.

1.3.1 The general research statement

In one sentence, the general research question of this thesis could be:

RESEARCH QUESTION: What conditions favor the emergence of successful community norms and culture when the main form of interaction with the social group happens through the manipulation of a digital common?

I will now try to elaborate on this high-level question; later I will explain how it can be operationalized in a way that can motivate the present dissertation.

First and foremost, we should clarify what we mean by “successful community norms and culture”. The objective of this research is not to ascertain whether a collaborative project might be able to produce high-quality objects, but rather to understand what makes its participants collaborate with each other in a context that is potentially unsuited to cooperation. The question of whether Wikipedia is producing a high-quality encyclopedia – up to the point of making expertise worthless – is a deceiving one. Similarly, for FOSS development, we asserted that the model itself is not a guarantee for producing bug-proof software, but that nonetheless FOSS is interesting because it has certain characteristics that make it a valuable alternative to the proprietary model. As we saw with the case of the Nature study (Giles [2005]), quality is hard to define, let alone gauge,

and it depends on the context. This is problematic for us, as we need criteria with which to measure success for a peer production community.

The degree of participation of users, on the other hand, is something easily measurable and thus perfectly suited to determine if a project “works” or not. We can thus formulate hypotheses about user participation, build models based on these hypotheses, and verify whether such models are compatible with empirical data from existing mass collaboration systems. Thus we are going to measure whether a peer production system is successful in terms of user participation.

We still need an operational definition of user participation that can be applied to a wide range of peer production systems. What is distinctive about peer production, as opposed to other forms of intellectual work, is that information artifacts are in common. They are available for consumption and evaluation to other users, in the case of content-sharing sites; but also for further manipulation, e.g. with wikis. Thus we can broadly define participation as the act of interacting with those artifacts that constitute the virtual common of the site. The activity lifespan of a user will therefore be the period during which we observe him contributing to the common.⁵ This information is easily accessible from the logs of user activity that the collaboration platform stores along the actions of the user, e.g. meta-data attached to each revision of a wiki page.

A high degree of user participation is therefore going to be, for this research, the condition of success of a community. What about the cultural traits needed to establish it? Users need not hold the same set of ideas, objectives, norms, or behaviors when they join a community of peer production. The main hypothesis of the present research is that interacting with shared content can bring individuals to be more alike with the rest of the community of users – as far as certain characteristics related to the production process and the general community life are considered. We generally refer to this as community norms. If these community norms are perceived to hold, all else being equal, a user will keep on participating to the community.

A good example of such a community norm is the NPOV rule of Wikipedia. Since, as we have seen, it is highly normative to write according to a neutral point of view, a new user will have to adhere to it, or his contributions will be rejected by the community. The example of NPOV can be generalized. In general users will have different standards about what to contribute and how. For example they might want to contribute content that other users do not find acceptable, or the quality of their contribution may not be up to the community

⁵We should note at this point that forms of passive participation e.g. lurking are not ruled out by this definition. However, collecting data on lurking users can be difficult. We discuss this chapter 3.

standards.

The last point that needs to be elucidated in our general research question is why we speak of the manipulation of a common pool of information. At this point the reader should have clearly understood that peer production is about building artifacts using a digital platform – encyclopedic entries, source code, product ratings, etc. In these contexts, social influence originates from the actions of other peers on the content that is collaboratively built. In a traditional setting the discussion about a community – its history, its rules, etc. – and the community life itself would be clearly separated. This is not the case anymore with online communities, because the collaboration medium is capable of archiving discussions, bug reports, frequently-asked questions, etc. Reagle [2007b] notes that this is especially true for wikis, where the discussion about what rules and policies to follow is part of the contents of the wiki itself.

1.3.2 Operational research question

Having set out the general goal of this dissertation, we can elaborate on the operational aspects of this dissertation:

1. The first set of questions is about empirical patterns of user participation; this is a crucial step, needed to establish those assumptions that will be later useful during the modeling phase. In particular, we are interested in answering the following questions:
 - (a) What distribution best describes user activity lifespan?
 - (b) Are there quantitatively similar patterns of participation across different communities?
 - (c) Can we infer what stage of development a community is at, by looking at the participation patterns?
 - (d) Is user activity homogeneous over time?
2. The second set of questions is related to modeling peer-production:
 - (a) Is user participation an emergent phenomenon of the dynamics of user contributions?
 - (b) Is content popularity a relevant factor in determining participation?
 - (c) Is initial motivation a relevant factor for explaining user participation?

- (d) Is it possible to compare the prediction of a computational model of user participation against empirical data?

QUESTION 1a: What distribution best describes user activity lifespan?

Answering this question is the first mandatory step to be able to perform any further research on user participation. However, the literature is contradictory on this issue. Leskovec et al. [2008] have reported that the distribution of activity lifespan in blogs follows the exponential distribution; Guo et al. [2009] have found instead that the age of object in similar contributory social networks is instead bimodal; Grabowski and Kosiński [2010] developed a model of user participation whose prediction is that participation lifespan follows a power-law decay, unfortunately without testing it against empirical data on lifespan. It would be thus interesting to understand if observations were denoted by the existence of a single characteristic scale, more than one, or none at all (i.e. a scale-free distribution), and what functional form of the distribution of user lifespan is best supported by the data.

QUESTION 1b: Are there quantitatively similar patterns of participation across different communities?

Is each peer production community a story on its own, or are there similarities across different communities? We are dealing with quantitative similarities. Of course if we look at two different systems, for example a blogging community and a discussion group, we will see several differences due to the interface design, the type of interactions going on (writing personal diary entries versus written conversation), and the respective popularity. But what if we were able to control most of these factors? Does user participation depend on the details of the system we are analyzing, or does it only depend on the dynamics of the social interactions under study? This concept is known in physics as universality. Several social systems have universal features, for example elections (cf. Fortunato and Castellano [2007]) or the distribution of scientific citations (cf. Radicchi et al. [2008]).

QUESTION 1c: Can we infer what stage of development a community is at, by looking at the participation patterns?

Because of their decentralized nature, peer production communities are under constant change. Can we infer the patterns of development of a peer production community from observation of the activity lifespan of its members?

QUESTION 1d: Is user activity homogeneous over time?

If we shift our focus from the level of the whole community to that of the individual contributor, how can we characterize individual activity? For example, are contributors similar in terms of contribution? Previous research, in general rules out this possibility: Guo et al. [2009] and Radicchi [2009] report a high level of heterogeneity in the level of contributions by users. Taking this into account, can we say anything about the general life cycle of the activity of an individual user? Is participation characterized by distinct phases of activity over time, for example?

QUESTION 2a Is user participation an emergent phenomenon of the dynamics of user contributions?

In other words, is it possible to characterize the macroscopic patterns of user activity in peer production systems using a microscopic model of user behavior? We have already elaborated on this before.

QUESTION 2b Is content popularity a relevant factor in determining participation?

Do the dynamics of information access affect the level of user participation? The way people consume and produce information is considered important in understanding how the general patterns of a peer production community unfold. This is also relevant for the topic of information quality, especially when contribution is open like in wikis. Indeed, the initial objection against Wikipedia was that, as the encyclopedia grows, vandalism would become more and more difficult to monitor; the degradation in the quality of the information would thus drive legitimate contributors away. In particular, when users compete for accessing a small subset of popular artifacts, we are interested in understanding if this will affect their patterns of user participation.

QUESTION 2c Is initial motivation a relevant factor for explaining user participation?

Is user participation explained only in terms of the strength of the initial motivation of a user? Or is it determined by the overall dynamics of interaction with contents? In other words, if users are initially more motivated, does this translate into a longer lifespan? And, more importantly, are long-term forms of participation only explained in terms of initial motivation?

In fact, our model is going to contain several tunable parameters. In order to assess the relevance of its various parameters we need to perform a systematic screening of all relevant factors. Only in this way we may hope to give a sensible answer to this and to the previous question.

QUESTION 2d Is it possible to compare the prediction of a computational model of user participation against empirical data?

This is a methodological question. Agent-based modeling is an attractive computational technique for the social sciences because it lets one perform experiments on what mechanisms may be responsible for given behaviors and collective phenomena. The mere ability to generate a plausible scenario, though, is not enough for claiming that a certain mechanism is responsible for a specific phenomenon, because often the parameters of a model can be tuned in a way to generate almost any possible response. In the natural sciences, the correct way to test a model is to compare it against data. Can we estimate the parameters that best approximate our data? Can we measure how well our model fits the data?

1.4 Main contributions of this dissertation

The contribution of this dissertation is twofold. The first is an empirical investigation of the lifespan of user participation, the evolution of user activity in Wikipedia, and a characterization of the life cycle of user activity. The second is a computational, agent-based model of peer production that connects social influence, bounded confidence, and user participation. A key result of this model is that the ease with which the process of social influence happens, the general level of tolerance towards changes of attitudes by users, is the most important factor in determining the overall patterns of user participation in a peer production community. Moreover, we address the methodological problem of the comparison of an agent-based model with empirical data; we develop an indirect inference technique to calibrate the parameters of our model using empirical data. This lets us estimate the parameters of the social influence for existing communities, and thus study the process of community formation at a quantitative level.

The following parts of this thesis have been published in peer-reviewed venues:

1. The paper “Empirical analysis of user participation in online communities: the case of Wikipedia” (cf. Ciampaglia and Vancheri [2010]) was presented at the 4th International AAAI Conference on Weblogs and Social Media (ICWSM’10), held in Washington D.C., on May, 23–26, 2010. This study covered part of the empirical part of this dissertation, in particular the investigation into the distribution of user lifespan.
2. The paper “A bounded confidence approach to understand user participation in peer production systems” (cf. Ciampaglia [2011]) was presented, as a full paper, at the Third International Conference on Social Informatics (SocInfo’11), held in Singapore on October, 6–8, 2011. That study covers part of the analysis of the model of peer production proposed in this dissertation.

In the next chapter we will review the major research areas in which this thesis is situated – social informatics, and the computational approach to the social sciences. In giving a bit of context about the wider research questions this thesis tries to address, we will argue that computer science has more connections with the social sciences than one would expect, and that therefore it is not so surprising to see computer scientists dabble in questions of a sociological nature, and, of course, the other way round.

Chapter 2

Literature survey

Every computer scientist is a failed psychologist.

Judea Pearl

This thesis is about user participation in commons-based peer production. As we saw in chapter 1 this is an important problem, since the diffusion of on-line communities of peer production has the potential to bring about a radical change in areas such as information technology, management, and governance. While this would be enough to motivate a study of peer production, I argue that it is not all. Online systems of peer production offer, from a scientific perspective, a unique opportunity to study collective human behavior in a simplified environment. This matters chiefly to two areas of inquiry. The first pertains to Informatics itself, and is the study of the social aspects of computing. The second area is that of social sciences at large: as Watts [2007] and others have noted, the introduction of computational methods for the analysis and modeling of collective social phenomena is bringing about a small revolution in this field.

Even though the research presented in this dissertation could easily belong to the area of social informatics, we should keep in mind that peer production is a highly multidisciplinary problem. It is thus useful to think of it in terms of connections between different disciplines, which is what I will strive to do in this chapter.

2.1 Structure of this chapter

The present research lies at the intersection of three major areas, so this chapter is going to be structured accordingly. The areas are: computer science, physics, and the social sciences.

The first section deals with the study of the social aspects of informatics and computing (section 2.2). Social informatics predates the study of social aspects of computing even though there is little discontinuity between the two denominations. After an initial historical excursus on the discipline of social informatics (subsection 2.2.1), I will introduce the concept of socio-technical systems (or *STS*, see 2.2.1), and explain how the World-wide Web can be thought of as an example of an *STS*, and finally highlight its importance for software engineering because it connects with the study of software development teams (2.2.1).

In recent years, researchers have started to look at the computational capabilities of groups of people. Collaborative filtering is perhaps the earliest area that marked this difference with previous social informatics research (2.2.2). The earliest applications of collaborative filtering techniques occurred mostly in corporate or institutional settings. It is with the so-called Web 2.0 revolution that we see the rise of large collaboration communities (2.2.2). Peer production, in a sense, begins here. With it, research moved to characterize models and theories of social computing (2.2.2).

We then move to the second major area of research, the area of socio-physics (2.3). In recent years, the hard sciences have been more and more interested in studying human behavior and society. This might seem rather novel, but I will show that this research agenda is several centuries old (2.3.1). In reviewing this area we start dealing with the general-purpose theoretical paradigm of complex networks (2.3.2), which is proving very effective in describing generic complex structures – social networks being a classic example.

Socio-physics is important to the present dissertation because of opinion dynamics (2.3.3). In fact, our peer production model can be regarded as an opinion dynamics model over a dynamic bipartite structure. We are specifically interested in the class of models of continuous opinion dynamics under bounded confidence (see 2.3.3). Bounded confidence is a concept borrowed from social psychology, and we will talk about it more in the third section of this chapter (2.4.2). We also cover the other models, in particular the other major class of discrete opinion models (2.3.3); we close looking at empirical investigations of opinion dynamics (2.3.3), namely elections and collective phenomena on the Internet such as crowdsourcing. Crowdsourcing will be discussed more in more

detail when we come to talk about the social psychological effects behind it (2.4.2).

Another important aspect of socio-physics is related to the study of human dynamics (2.3.4). Many quantitative aspects of human behavior, such as communication and information processing are now accessible thanks to the availability of large datasets of online interactions, such as emails or text messages. Because our analysis of user activity lifespan is interested in quantifying some temporal aspects of Wikipedia editors, this dissertation relates to several studies in that context.

The third major section is related to the socio-economics of online peer production communities (2.4). The most pressing research problem is the incentives structure of people participants to peer production communities (2.4.1). Especially relevant to this research is the social psychology of user participation (2.4.2). In this context I will specifically discuss the connection with opinion dynamics models through the case study of crowdsourcing (see 2.4.2).

Besides these three major sections, we also need to give some attention to our particular case study – Wikipedia, the free online encyclopedia. This is the reason for the inclusion of a fourth section (2.5) specifically devoted to the research on Wikipedia. Of course we do not cover all the research on the topic, but only those works that have something to do with our research. In particular, after discussing Wikipedia as a case study of peer production (2.5.1), we review the works on the governance of such a large community (2.5.2, and on the quality of the encyclopedic entries (2.5.3).

2.2 Social informatics and social computing

2.2.1 Social Informatics

As already said, the research presented in this dissertation could belong to the field of social informatics. In the words of Kling [2007, p. 210]:

One key idea of social informatics research is that the “social context” of information technology development plays a significant role in influencing the ways that people use information and technologies, and thus influences their consequences for work, organizations, and other social relationships.

In this view, the social context mainly refers to the “particular incentive systems for using, organizing and sharing information at work” (*ibidem*.) From this

passage we can already note two things: the first is that what is of interest here is primarily the development of information technology; in particular, researchers want to understand how the social environment in which a piece of information technology is developed affects its structure and its functions (cf. Dutton [1997]; Huff and Finholt [1994]; Kiesler [1997]; Smith and Kollock [1998]). Similarly, this discipline argues that the organizational context in which a technology is used is relevant for understanding its success or its demise (cf. Lamb [1996]). Traditionally, social informatics has been more interested with understanding how, and under what conditions, the computerization of organizations and institutions would lead to improvements in productivity (cf. Robey [1997]; DeSanctis and Fulk [1999]).¹ The focus on organizations and not on the broader society was mainly due to historical reasons: computerization, at least in developed countries, happened first within companies and other large organizations, and only later involved the rest of society. This leads us to the concept of socio-technical systems (STS).

Socio-technical systems

Software engineers traditionally reason in terms of use cases, that is, idealized scenarios by which they describe the interactions between the system they are developing and the external world – users, administrators, and other systems (cf. Cockburn [2001]). What is important in a use case is that people are not so much identified by their social status, but by the role they happen to hold during that particular situation.

By contrast, a socio-technical system encompasses all the relevant external world, and tries to identify human actors by their social role, thus taking into account norms, organizational hierarchies, and so on. Historically, social informatics researchers were those studying the factors of success of digital libraries – why some information systems such as electronic journals would work while other, technically similar, would not (cf. Bishop and Starr [1996]). In this respect, it was an area of research that belonged to the computer-supported cooperative work (CSCW) field and was concerned with understanding behavioral factors (cf. Kling et al. [1998]).

Needless to say, the concept of STS has wider implications than just information technology. All forms of engineering, one could argue, have been influenced

¹This is not to say that social informatics was only about the effect of computerization in organizations. Other questions were more in line with the computer-supported collaboration work (CSCW) research, such as the development of trust in virtual environments, cf. Iacono and Weisband [1997].

by historical, economical, and sociological factors.²

An example: the Web

To describe what an STS is, and what type of problems can be framed under this concept, let us consider the Web. The Web was initially intended as an information repository for the high-energy particle physics community (cf. Gromov [2000]). Even before the Web was invented, access to information on the Internet was open in that the community of users was small enough for people to trust each other (cf. Abbate [1999]). As the technology became more and more popular, the Web became a communication infrastructure used to perform commercial transactions. New technologies, like the secure-socket layer (SSL) were introduced, to ensure the necessary security, as the medium was now used to exchange sensitive information – for example credit card numbers (cf. Rescorla [2001]). This exemplifies the idea that an STS has to adapt when the socio-economical context it is used in imposes new patterns of usage.

Predicting changes in STS-s is far from being a trivial problem. From this brief description it is clear that STS-s are examples of complex, inter-dependent systems; even though, in 1999 Kling [2007, reprinted] had already noted, when discussing the research of the 70s and 80s, that “*analytical failure of technological determinism is one of the interesting and durable findings from social informatics research*,” technological prediction is still very much a trendy exercise for non-academics nowadays. With progress in large-scale data collection things are starting to change, as Vespignani [2009] points out, but one should take into account that the prediction exercise in the social sciences takes a very different meaning than, for example, the natural and physical sciences.

Nonetheless, fringe academic circles routinely build on empirical laws of technological growth, such as Moore’s Law (cf. Schaller [1997]), to predict that technological progress will soon reach a faster-than-exponential growth, with major consequences for society and civilization. In contrast, social informatics scientists are interested in understanding under what conditions adoption of a technology will lead to a successful outcome or not. In this sense predictions are much more limited in scope, even though they often lend themselves to deeper interpretation.

²According to Kling [2007] the denomination of STS was first used in 50s by an English school of psychologists that were interested in understanding the well-being of workers in various production and manufacturing contexts.

Software development and organizations

The problem of adaptation to exogenous trends is not just a prerogative of communication media like the Web, though. Any piece of information technology exists in a social context, and the structure of different social groups, with their specific norms and behaviors, influences its evolution over time. Software engineering, in a sense, was intended specifically to solve this problem, a fact denoted by early empirical principles relating productivity and team size. In his seminal 1975 book, Brooks [1995, reprint] noted how “adding manpower to a late projects makes it later”. This is just a witty way to say, Raymond [1999] argues, that organizational complexity rises faster than the team size, in particular as $O(n^2)$, where n represents the team size. The idea that organization influences software quality is starting to be validated empirically: Nagappan et al. [2008] found for example that organizational indicators are the best predictors for failure-proneness of software components.

Based on these considerations, the way open source teams are organized has been often pointed to as a key advantage of its software development model. According to Raymond [1999] this happens because of a subdivision of concerns. The core of a development team is often composed by only a handful of individual developers. This is the level where major design decisions are taken, and the quadratic scaling in communication complexity due to Brooks’ law does not create problems thanks to small sizes. Testing, on the other hand, is performed by the community at large. This activity can be easily split up into independent tasks, hence requiring little or no communication among testers.

Another, perhaps more intriguing, principle is the one due to Conway [1968], which states that the structure of a piece of software reflects the structure of the organization that has developed it:

[...] organizations which design systems [...] are constrained to produce designs which are copies of the communication structures of these organizations.

The usual Raymond [1999] gives a perfect example of this: “If you have four groups working on a compiler, you’ll get a 4-pass compiler.” There is much consensus, at least at an empirical level, on the validity of Conway’s law (cf. Herbsleb and Grinter [1999b,a]; Bowman and Holt [1998]; Amrita and van Hillegersberg [2008]), but how to characterize the structure of a piece of software? Bird et al. [2008] propose to look at the communication structure, in particular mining the mailing list archives of a project.

2.2.2 Social computing

There is also another aspect of peer production that computer scientists find increasingly intriguing: the computational capabilities of groups of people. Technology commentators coined several buzzwords for this phenomenon. Surowiecki [2004] first popularized the concept of “wisdom of the crowds”, O’Reilly [2005] stressed the importance of the technological change by coining the term “Web 2.0”, and Tapscott and Williams [2006] argued that mass collaboration is even bound to change global economies. The study of social computing is different from that of social informatics because it is motivated by practical problems, e.g. how to recommend a relevant movie given the tastes of a user. In fact, if we compare the earliest social computing works with the contemporary research published in social informatics venues, such as Stodolsky [1995], we see that there is little interest in understanding the social context of a user, or even in acknowledging that people may form their tastes and opinions based on those of their peers. Only recently theories of how social influence affects the formation of user tastes and how computing applications can take advantage of this have been set forth. We touch on this point briefly in the present section and will devote sections 2.3 and 2.4 to discuss this problem more in detail.

Collaborative filtering

The first social computing applications arose as an alternative to the problem of organizing the contents of large information systems – a problem of retrieval of information.

Traditionally this problem is solved by looking at the specific properties of the contents of a system. This content-based approach represents the digital contents of any document, be it an email, a movie, or a song, using a suitable mathematical formalism, and then selects, from a collection of novel documents, those that are most similar, in a mathematical sense, to what a user likes (recommendation), or that meet a certain set of criteria (relevance filtering).

In contrast, collaborative filtering applications are motivated by the observation that these cognitive tasks are best solved by people, and that therefore what is needed is just a way to aggregate the evaluations of a large crowd of users. Early collaborative filtering applications were developed in the mid-90s for types of digital contents such as electronic mails (cf. Goldberg et al. [1992]), news (cf. Resnick et al. [1994]; Hill and Terveen [1996]; Konstan et al. [1997]), videos (cf. Hill et al. [1995]), and music (cf. Shardanand and Maes [1995]).

Of course once we relied on the judgments of users we exposed to the dan-

ger of malicious users inserting bad evaluations or simply junk, so in the years that followed the problem of trust and reputation became increasingly studied, examples are: Resnick et al. [2000]; Dellarocas [2003]; Golbeck [2005]; Adler and Alfaro [2007].

The field of recommender systems evolved considerably after this first pre-2001 wave of works, in particular see Burke [2002]; Herlocker et al. [2004]; Adomavicius and Tuzhilin [2005], but most of the efforts were towards the engineering component. Comparably, little attention was devoted to understand possible biases due to social influence and, in general, the dynamics of communication. See Sabater and Sierra [2005] for a review on recent developments in the field.

Web 2.0 and the rise of collaboration communities

One of the earliest references for the term “social computing” is by Schuler [1994] as, “describing any type of computing application in which software serves as an intermediary or a focus for a social relation,” it is only after O’Reilly [2005] had theorized the “Web 2.0” paradigm that it earned the status of “trend”, as evidenced by Wang et al. [2007], who discuss the shift from the old social informatics research. Parameswaran and Whinston [2007] addresses some of the research issues in the field, in particular those related to building social capital. The big difference with early collaborative filtering application is probably the rise of mass collaboration communities such as blogs, wikis, Q&A sites, and so on.

There are probably too many examples of mass collaboration communities to be listed. For a recent taxonomy of social computing applications, see Quinn and Bederson [2011]. Here I would like to cite two examples of how groups can solve real computational problems. The first is the Mechanical Turk (MTurk, in short), a marketplace for labor where users can post so-called human intelligence tasks (HIT) and other users can offer their labor. According to Mason and Suri [2011] the MTurk is an attractive option for behavioral scientists who wish to perform experiments; Snow et al. [2008] also reported a positive experience with computational linguistics tasks.

The main incentive for workers in the MTurk is direct remuneration. But people can be motivated to offer their labor for different incentives, such as entertainment and fun, and outcomes need not have poorer quality. A very recent result is the one by Khatib et al. [2011], who showed how players of the protein folding game Foldit managed to identify the best configuration of a retroviral protein connected to the AIDS disease. These are not isolated results

anymore. Benkler [2002] also cited the NASA Clickworkers program as another crowdsourcing application where people helped to analyze data for scientific experiments in the natural sciences.

Models and theories of social computing

Whereas the application of social computing ideas is already at an advanced stage, the theoretical understanding of it is still in its infancy. Models have been proposed for platforms that address specific problems, like the work of Lerman [2007a,b] on the social information filtering community Digg. Golder and Huberman [2006] explored tagging applications, Wilkinson and Huberman [2007] model the accretion of the number of contributions to Wikipedia pages in order to estimate the correlation between information quality and the amount of collaboration going on inside Wikipedia. Crandall et al. [2008] elaborate on the model of Holme and Newman [2006] to model a feedback effect between social influence and user interests – an observation of potential interests for the problem of social recommendation. Recommender systems have also returned into play as a means to help oversight and task allocation in large communities (cf. Cosley et al. [2005, 2007]). Task allocation and self-selection has been explored by Li and Hitt [2008]. Wu and Huberman [2008] analyze the dynamics of polarization of ratings for movies and books, and Lorenz [2009] showed how an averaging process can explain the discretized statistics of movie ratings on the popular movie database IMDB. Regarding user-contributory websites, some stochastic models have been proposed (Hogg and Lerman [2009]; Hogg and Szabo [2009]). Wilkinson [2008] models user contributions in three existing peer production communities as a preferential attachment process with aging. A model of online communities by Grabowski and Kosiński [2010] features a power-law prediction for the lifespan of users, although Guo et al. [2009] and Ciampaglia and Vancheri [2010] find evidence for a bimodal distribution. Wu and Huberman [2008] analyze data from the news aggregator Digg and found that the collective attention towards news items decays with a characteristic time scale.

Is there any theoretical framework of social computing? von Ahn et al. [2005] formalize the concept of “Games with a purpose” (cf. von Ahn and Dabbish [2008] for a general introduction). Chevaire et al. [2007] give an introduction to several computational aspects of social choice, for example the problem of finding fair rules for an election, a problem that stretches back to the seminal work of Bartholdi et al. [1989].

2.3 The dynamics of collective social phenomena

2.3.1 Physics and society

The idea that human behavior follows a universal “code of nature” has a long philosophical tradition, stretching back to the writings of Adam Smith and Thomas Hobbes. The idea that the laws of collective human behavior can be stated in a quantitative fashion is not much earlier either. A “physics of society” was in fact the objective of Nineteenth Century statistician Adolphe Quet  let, who extensively studied the context in which the law of errors could lead to a description of society in terms of “average man”, and sociology pioneer August Comte who, in his *Course on Positive Philosophy* (1830–1842) wrote:

Now that the human mind has grasped celestial and terrestrial physics, mechanical and chemical, organic physics, both vegetable and animal, there remains one science, to fill up the series of sciences or observation–social physics. This is what men have now most need of; and this it is the principal aim of the present work to establish.

However, this connection is not unidirectional. In fact, as Ball [2002] notes, James Clerk Maxwell himself was aware of the work of Quet  let when he first set out to give a foundation to statistical mechanics.

Of course the atoms in a gas are quite different from the individuals in a society, and such an impediment has always led people to regard the connection between physics, statistics, and sociology as a mere historical curiosity. Things started to change in the second part of the 20th century, when the study of human behavior found a first formalization by means of Game Theory by Von Neumann and Morgenstern [1944].

Schelling [1971] is famous for being one of the first to study a collective phenomenon – the dynamics of segregation resulting from the interactions of individuals from different groups; at the same time Weidlich [1971] was probably the first physicist to model the dynamics of collective opinion formation by studying polarization in groups. His framework, called socio-dynamics, is related to the approach of synergetics, introduced by Haken [1978].

In later years, several models from statistical physics have been studied to understand social phenomenology; the Voter model and Ising’s model are perhaps the most famous (cf. Castellano et al. [2009] and references therein). In these models a network represents the social structure between agents and each agent has the choice between a binary set of opinions. In fact, ideas of graph

theory have proved to be instrumental for the development of the study of collective social systems (cf. Watts [2004, 2007]; Vespignani [2009]) therefore we will start the treatment of the literature of this field, which will roughly follow the taxonomy given in the review by Castellano et al. [2009] before, with an excursus on the field of complex networks.

2.3.2 Complex Networks

Before getting into the rich topic of complex networks, it should be noted that this thesis is not directly involved into the analysis of complex social networks from online communities of peer production. We make use of some tools from this field, namely preferential attachment, but only in a very limited fashion. Nonetheless, several other concepts of socio-physics require the understanding of key concepts from this field, therefore we review it here briefly.

Graphs are old concepts in mathematics and computer science. The statistical study of generative mechanisms of random graphs, in particular, was started by Erdős and Rényi [1959], who first studied a model of purely random connections connecting the nodes of a graph of prescribed size – hence the name Erdős-Rényi (ER), or Poisson, random graph. But it is only with the publication of the influential work by Watts and Strogatz [1998] that the theory of networks (new disciplines always forge new names for old concepts) found concrete application in the modeling of real systems. In that paper, the authors studied a class of networks called “small worlds”, which lies in between two extremes: the perfect regularity of lattice structures, and the total randomness of ER graphs. Small worlds are peculiar because, in contrast with regular lattices, they have a small average path length and, unlike Poisson random networks, nodes are clustered. Both these properties have made the class of small world networks a model for understanding many real network structures, such as social networks, the electrical grid, or biological neural networks.

Another important connection between statistical physics and graphs was shown by Barabási and Albert [1999], who proposed a preferential attachment model (PA) as a generative mechanism for a large number of existing network structures.³ The idea of PA is simple: as new nodes are created, they are connected with existing ones with probability proportional to the number of connections the existing nodes have. They showed that PA leads to scale-free network

³Preferential attachment had been already studied, under different names and in different contexts, by Undy Yule and by Herbert Simon, but not as a network growth model. de Solla Price [1965] is credited as the first to apply the idea of PA to a network, in particular the network of scientific citations.

structures, in that the distribution of degrees follows a power-law distribution.

In the years that followed complex networks were extensively studied (cf. Albert and Barabási [2002]; Newman [2003b]; Dorogovtsev and Mendes [2003]; Durrett [2007] for introductory texts). Network growth was one of the main topics. The generalization of the preferential attachment rule was studied by Krapivsky et al. [2000]. A constant term, the initial attractiveness of nodes, was added to PA by Dorogovtsev et al. [2000]. Aging effects were taken into account by adding the “death” of nodes by Dorogovtsev and Mendes [2000]; Lehmann et al. [2005]; Lambiotte [2007]. Preferential attachment is an example of a reinforcement process; for a recent survey of reinforcement processes for network growth, see Pemantle [2007].

Simple generalizations of the ER random graph model with arbitrary degree distribution were studied by Newman et al. [2002] these models are solvable in the limit of an infinite network size.

As we have already seen in the work of Watts and Strogatz [1998], another characteristic that is not captured by the ER model is the presence of a structure, or hierarchy, within empirical networks. For example, tightly connected subgroups may be connected together by a few “long-range” edges. In the already-cited “small-world” paper of Watts and Strogatz [1998] this characteristic is measured in terms of the clustering coefficient, which measures how much each of the neighbors of a node are connected with each other.

In fact, this property of the connectivity of nodes is related to another property called assortative mixing. Assortativity is the general tendency of nodes in real-life networks (e.g. people, websites, animals) to form connections with nodes sharing similar features. Examples of assortative mixing arise in contexts: in social network analysis this is in fact a well-known phenomenon and goes under the name of homophily (McPherson et al. [2001]). Assortativity can be thus responsible for the existence of a community structure within a network. In the extreme case of a maximal assortativity, a network is in fact split in multiple sub-communities of similar nodes, with few or no edge connecting nodes with different features (Newman [2003a]). The problem of discovering the community structure in a network is a very active area of research at present, and is based essentially on these ideas (Newman and Girvan [2004]). A recent review of Fortunato [2010] covers this rich multidisciplinary field.

Of course one can also focus on specific building structures of a network, which are called network motifs Milo et al. [2002]. For example, a feature of social networks of friendship is called triadic closure, that is the fact that friends of a person are likely to be friends themselves. This leads to an over-expression of triangles in the network topology. In general it is important to understand

what the characteristics of social networks are, because social influence might be a key factor in shaping them Kumpula et al. [2009].

Another important application for the study of complex networks is the study of epidemics (cf. Moore and Newman [2000]; Newman [2002]).

From the point of view of peer production, network models have been applied for studying: the Internet graph Faloutsos et al. [1999], the blogosphere Adar et al. [2004]; Leskovec et al. [2008]; online groups Backstrom et al. [2006]; collaborations and controversies in Wikipedia Brandes and Lerner [2007, 2008]; Brandes et al. [2009b]; the hyperlink structure of Wikipedia Capocci et al. [2006, 2008]; Zlatić et al. [2006]; the World-wide Web Huberman and Adamic [1999]; Broder et al. [2000]; scientific citations Newman [2001]; and mobile call networks Seshadri et al. [2008].

2.3.3 Opinion dynamics

The problem of finding a consensus in a group of people over a certain matter of discussion is a fundamental problem of group decision-making and cooperation and as such has been studied by sociologists, psychologists, economists, and political scientists (Davis [1973]; Laughlin [1980]; Latané [1981]; Stasser et al. [1989]; Friedkin and Johnsen [1990]; Latané [1996]; Witte and Davis [1996]; Friedkin and Johnsen [1999]). Consensus has been studied at different levels, starting from how small groups agree on a decision, up to the level of a society, where we are interested in understanding how public opinion forms, and how political parties come about, for example.⁴ Thus group formation is also a problem related to opinion dynamics and group decision making.

The first question one might ask is: does it make sense to treat the dynamics of the opinions of people mathematically? Of course, even in small groups, consensus may not be easy to describe: opinions are in general multifaceted and ineffable. It is in human nature, after all. On the other hand, it is not uncommon to face very simple “discrete” choices, for example coffee vs tea, Emacs vs Vim, etc. Therefore, even though in general opinions are objectively difficult to measure, there are many other cases in which describing an opinion using one or more numbers is perfectly legitimate; looking at such simplified settings may, in turn, open the way to a mathematical treatment of the general dynamics of opinions in a social group. So, as Weidlich [1971] points out, looking at simple cases might give us some insight into the case of more complicated opinion

⁴The problem of consensus can be also formalized and studied in an abstract setting, where instead of people we have a distributed system and its components have to agree over a certain course of actions. This is one of the fundamental problems of distributed systems research.

structures.

Discrete opinion dynamics

As with complex networks, statistical physics can provide the right framework for understanding the dynamics of consensus, polarization, and fragmentation of large groups of interacting agents (cf. Castellano et al. [2009], and references therein). The first works on opinion dynamics are those of Weidlich [1971] on polarization in the framework of socio-dynamics, (cf. Weidlich [2000]). Following these, we have the voter model (Clifford and Sudbury [1973]; Holley and Liggett [1975]), and the works of Galam et al. [1982] on majority rule. These models are directly inspired to statistical physics models. They feature a discrete opinion, usually a simple binary choice, in the same way as we can think of an electron spinning in two distinct ways. Even though we do not deal with discrete models of opinion dynamics in the context of this research, it is worth mentioning that they have been studied extensively, first in the field of probability theory (Cox and Griffeath [1986]; Liggett [1985]), and later as versions of the Ising model de Oliveira et al. [1993]; Scheucher and Spohn [1988]. Of course the role of a realistic social structure has been taken into account, by considering network topologies different from the regular lattice that was featured in their original formulations (Castellano et al. [2003]; Wu et al. [2004]; Wu and Huberman [2004]; Castellano [2005]; Sood and Redner [2005]; Michard and Bouchaud [2005]). The excellent survey article by Castellano et al. [2009] analyzes the work on discrete models of opinion dynamics extensively.

In the context of this thesis we are more interested in a specific class of opinion dynamics models instead. Moreover, rather than discrete variables, these models use continuous variables. This allows for more nuanced definitions of “opinions”, as we can take into account a range, e.g. the political spectrum of modern democracies that goes from Left to Right Deffuant et al. [2001]; Castellano et al. [2009]. These models are useful to study how consensus develops in groups of people, committees, political parties, etc. In particular, models with continuous opinions have been developed to study an aspect of human communication that in the field of social psychology is known as *bounded confidence* (BC). We will see the bounded confidence rule later, in the context of two theories of social psychology, namely the theory of Self-Categorization and the theory of Social Judgment. The idea here is that an exchange between people can make their views more alike only if their original opinions on the matter were not too distant. How distant, at the most, two opinions may be is given by the so-called confidence parameter $\varepsilon > 0$.

Continuous opinion dynamics

The two most famous models are the model by Deffuant et al. [2001] and that by Hegselmann and Krause [2002]. Discrete models of opinion dynamics are interesting because of their analogy with spin models and for the possibility of taking into account different kinds of network topology. They are motivated by the fact that individual agents change their state after interaction with their neighbors. In this way, the network topology reproduces the underlying social structure in a group of interacting agents.

Although attractive for their connection with spin models, which opens up the possibility of using the machinery of statistical physics to analyze their behavior, spin models (as well as majority-rule models and similar) were not satisfying for many researchers because they were not taking into account several important aspects of human communication, and because the discrete assumption seemed too simplistic for those cases where opinions are not restricted to a finite set of well-separated choices.

More or less at the same time of the early uses of spin models for group consensus, the problem of modeling the salient features of communication within groups of experts was already being studied by mathematicians and statisticians (Stone [1961]; Chatterjee and Seneta [1977]; Cohen et al. [1986]); this is indeed a classic problem in economics too (Visser and Swank [2007]), and these studies eventually gave rise to the famous Delphi technique DeGroot [1974]; Linstone and Turoff [1975] developed at the RAND corporation. Experts in a committee usually have to give a quantitative evaluation (hence the continuous assumption) and in general no single committee members possess perfect information. This means that individuals have to adjust their opinion according to the signals they receive from others. This leads to the concept of bounded confidence, that is, the fact that there will be an adjustment only when the two original positions were not too dissimilar.

Bounded confidence is taken into account in both the Deffuant and the Hegselmann-Krause model by introducing an averaging dynamics; the difference between the two models is that the first features pairwise interactions while in the latter adjustments happen within finite-size groups. The Deffuant model has received somehow more attention in literature: Weisbuch et al. [2003] studied the role of a heterogeneous population of agents, Fortunato [2004] found that the threshold for complete consensus of $\varepsilon > 0.5$ is universal. Lorenz [2007b] formulated both models in terms of averaging of row-wise stochastic matrices and studied the phase diagram of consensus of both models. Several modifications have been proposed to model other aspects of consensus in groups, such

as the presence of extremists (Deffuant et al. [2002]; Laguna et al. [2004]); leaders and external propaganda (Carletti et al. [2006]); spontaneous drift of opinions (Ben-Naim [2005]).

A discretized version was proposed by Stauffer et al. [2004], while Martins [2009] recast the BC rule in terms of Bayesian updating; Lorenz and Urbig [2007] studied how the communication rule the BC agents use affects the likelihood of consensus; Carletti et al. [2008] studied a version of the Deffuant model where the population of agents was subjected to birth and death process; also, Carletti et al. [2010] studied a process of network formation induced by the mixing rule of the Deffuant model. Hegselmann and Krause [2006] studied the case when agents have to agree on a true value and how population heterogeneity affects this. A different version of the BC rule that better reflects the ideas of self-categorization, was studied by Salzarulo [2006]. Bounded confidence has been studied also in the context of individualization theory by Mäs et al. [2010]. Lorenz [2007a] surveyed the field of continuous opinion dynamics under BC.

Other models of opinion dynamics

Of course the previous works do not cover the whole field of opinion dynamics. Without claiming to be exhaustive, we should also cite the social impact model (Lewenstein et al. [1992]; Holyst et al. [2000]), we also cited the majority rule model (Galam [1986, 2002, 2005]). A different but related field of study is the one related to the emotional response of large crowds of people. Different from opinions, emotions are denoted by other characteristics, like their saliency, and therefore obey different rules. Gonzalez-Bailon et al. [2010]; Chmiel and Hołyst [2010]

Empirical investigations of opinion dynamics

The works we have seen before deal only with the theoretical development of the study of the dynamics of opinions; Sobkowicz [2009] argues that neglecting empirical aspects is a problematic aspect for the whole field, because of the weak connection with an empirical phenomenology. As we said before, it might be difficult to measure the opinion of somebody in certain contexts, for example politics, whereas in other contexts such data might be readily available. Nonetheless, there are several ways to link the insight provided by models of opinion dynamics with the real world. Filho et al. [1999] were the first to analyze data from elections in Brazil; later Fortunato and Castellano [2007] showed that the distribution of votes candidates attain in proportional elec-

tions is universal across several countries and years, and propose a branching model of opinion dynamics to account for the empirical data. Other models have been proposed such as the one by Bernardes et al. [2002]; Travieso and da Fontoura Costa [2006].

Michard and Bouchaud [2005] analyze, on the other hand, so-called collective swings, that is, situations where imitation and social pressure can give rise to herding effects, using a Random Field Ising model. In particular, they look at data on birth rates in different countries, sales of cell phones, and the reduction in clapping during a concert.

In other cases the “opinions” of people might be directly measurable, as in the case of product ratings and movie reviews, which are provided by rating communities such as the Internet Movie Database (IMDB). Lorenz [2009] showed that the binned distribution of the number of “stars” a movie receives on IMDB can be fitted to a confined Levy distribution, and notes that such a distribution arises in the context of an averaging process, hence again a connection with a compromise dynamics reminiscent of the Deffuant model. Wu and Huberman [2008] analyzed book reviews on Amazon and movie ratings on IMDB and noted that self-selection tends to mitigate the development of extremism in product reviews.

Opinion dynamics is also relevant to understand to what extent the “wisdom of crowds” effect works. In a recent experimental study Lorenz et al. [2011] found that even a small amount of social influence can make groups grow overconfident in simple tasks of prediction.

Co-evolution of state and topology

So far the models we have seen assume that the underlying social structure in a group, which is modeled by a network, is fixed over time. This assumption can be accepted for the sake of simplicity, and in some cases, for example with committees of experts, it is not even problematic to justify it, but in general it is somewhat limiting. Moreover, as with any other network, the development of a social network is also affected by its function (cf. Newman [2003b]). Social networks, in the end, are just the product of human interactions, and we know that humans tend to forge their social connection following basic principles. One of them is homophily McPherson et al. [2001], that is, the idea that “birds of feather flock together” (cf. Lazarsfeld and Merton [1954]).

The idea that certain features correlate between adjacent nodes is called assortativity (cf. Newman [2003a]; Zhou and Mondragón [2004]; Boccaletti et al. [2006]; Costa et al. [2007]), and can have profound implications on the result-

ing topology: high assortativity can lead, for example, to breaking a network into sub-networks, since every node will be connected only with those that are similar to it; assortativity is also linked to the problem of extracting the structure of communities in a network, cf. Newman and Girvan [2004]. On the other hand, models opinion dynamics are motivated by the idea of determining how the social structure affects the characteristics of an individual.

The interplay between social influence and link selection in social networks has been the subject of research by several authors. In particular: Holme and Newman [2006]; Gil and Zanette [2006]; Stauffer et al. [2006]; Nardini et al. [2008]; Kozma and Barrat [2008b,a]; Iñiguez et al. [2009]; Carletti et al. [2010]. This field of research is interested in understanding the dynamics of co-evolution of state and topology in social groups. Holme and Newman [2006] studied a phase transition in the consensus state of a simple model of an adaptive network with a dynamic of opinion formation (social influence) and a link rewiring mechanism (link selection).

From an empirical standpoint, Crandall et al. [2008] has evidenced how the interplay between influence and selection plays an important role in peer production systems. They showed the existence of a feedback effect between the two, and highlighted the implications for the design of recommender systems. Aiello et al. [2010] studied the phenomena of link creation and alignment of profiles in the book sharing community Anobii.

2.3.4 Human dynamics

Whereas opinion dynamics is interested in modeling the problems of consensus, agreement, negotiation, and prediction in social groups, the field of human dynamics is interested in understanding wider human behavior at a quantitative level, again using ideas from statistical physics, and taking advantage of large datasets of human activity, often taken from the Web (Baldi et al. [2003]), or from other technological networks (Vespignani [2009]).

Early works deal in particular with human communication, especially emails (Ebel et al. [2002]; Johansen [2004]), printing jobs Harder and Paczuski [2006], and later into the problem of the statistical identification of the distribution of inter-event times in general human communication (Oliveira and Barabási [2005]; Barabási [2005]; Dezsö et al. [2006]; Vázquez et al. [2006]; Gonçalves and Ramasco [2008]). These works found a scale-free behavior in the response times of human correspondence spanning several orders of magnitude. However, the topic is currently under debate because Poissonian models that take into account circadian rhythms and weekly trends have been shown to fit the

data better than power-laws (cf. Malmgren et al. [2008, 2009]).

The study of Web analytics is also related to this field. Datasets that have been investigated, in particular, are those on human browsing on the Internet Huberman et al. [1998]; Johansen and Sornette [2000]; Brewington and Cybenko [2000]; Wainwright [2003]. In the context of peer production, the inter-event time distribution between two consecutive actions on Wikipedia and other peer production systems has been found to decay according to a power-law Radicchi [2009]. Similar ideas have been used to model the cascades of edits to Wikipedia pages made during editing conflicts (Sumi et al. [2011]).

2.4 The socio-economics of online communities

2.4.1 Incentives to contribution

Peer production is puzzling because it challenges the widely held assumption that people would perform highly-skilled, intellectual work only in return of a remuneration (Hars and Ou [2001]; Lerner and Tirole [2002]). Open Source software developers for one are a good counterexample to this (Raymond [1999]).⁵ Is it a form of altruism, and why is it so puzzling?

In a sense, there is nothing extraordinary in altruistic behavior. Humans are pro-social beings after all. Nonetheless, as Open Source became popular it caught the attention of social scientists (together with the aforementioned Hars and Ou [2001]; Lerner and Tirole [2002], also Kollock [1999]; Benkler [2002]; McKenna and Green [2002]). What was puzzling – and novel – to them was the context in which this alleged altruistic behavior was happening. It was indeed unexpected to them to hear of software developers – not exactly an unskilled labor force – spending much of their time contributing to free software.

This is not to say that, according to economic theory, the only form of work people will do in exchange for anything other than a direct remuneration is manual labor. As a matter of fact, we know that there are several intellectual activities whose benefits are only extrinsically rewarding, that is, that are performed without any immediate direct reward (e.g. money) but with benefits

⁵Of course not all FOSS nor Free Software is contributed to by unpaid volunteers. Many companies pay some of their employees to work on FOSS projects they have a strategic interest upon. A good example is the Linux division at IBM Corporation, or Google. This is perfectly welcomed by the rest of the community, and does not hinder contributions from volunteers—as long as certain conditions are met, see Benkler [2002]. It has actually been argued that the opposite is true, i.e. by contributing for free to high-visibility FOSS projects, talented programmers have the chance to show their skills to potential employers.

that are going to pay off in the future (e.g. fame). Academics, for example, donate their time evaluating submissions to scientific journals. This activity has an opportunity cost related to the time spent doing it instead of doing more fruitful pursuits, so why do scientists do it (Benkler [2002])?

Economists construe it as a form of community service academics do in return for recognition from their fellow researchers. An academic with a higher reputation will stand more chances of finding a tenured position (Shatz [2004]). We are thus reconciled with the idea of people being rational agents and performing actions that will maximize their own utility, present or future. We will come back to scientific peer-reviewing as a good example of peer production that predates Internet communities (Spier [2002]).

However, there are no such things as tenured positions in the software industry, and since the beginning of FOSS it has been normal for employed programmers and sysadmins to contribute to projects in their spare time. Therefore economists, and sociologists alike, legitimately asked themselves why people would contribute to FOSS projects at all (Lerner and Tirole [2002]). What kind of motivation is there behind FOSS in particular, and peer production in general? Is it, in any way, similar to academic peer-review, e.g. a reputation-building activity? Could altruism and idealistic thinking explain all the incentive structure to participate in FOSS development? And do incentives work the same way across all projects?

The picture that comes from extensive surveying of FOSS contributors is that, rather than a single factor, the structure of incentives to participate is formed by an array of reasons (Kollock [1999]; Benkler [2002]; McKenna and Green [2002]; Wash and Rader [2007]; Rafaeli and Ariel [2008]; Anthony et al. [2009]; Brandes et al. [2009a]; Schroer and Hertel [2009]; Lampe et al. [2010]; Nov [2007]; Nov et al. [2010]; Fung et al. [2011]).⁶ Among these are purely sociological factors, like expected reciprocity, and economic factors, in particular a range of intrinsic or extrinsic benefits aside from simple remuneration. Reputation improvement is one of them for example, and in general all signaling incentives related to career-building concerns.⁷

⁶I am referring, in general, to “contributors” as opposed to “developers” because there are several important tasks, in the development of a software, that are not about software development *tout court*. For example documenting the software, doing translations, and managing common online resources (bug-tracker, mailing list, etc.) are all vital activities in any FOSS project. Although most analyses focus on code contributors only, many conclusions can generalize also to these other forms of contribution.

⁷Benkler [2002, p. 373, note 7] contends that extrinsic remuneration under the form of reputation building is a proper incentive for participating in FOSS by noting that two of the most famous FOSS projects do not publish the names of their contributors. In wikis, where attribution

From a sociological point of view, a characterization of FOSS that enjoys some popularity is that of a gift culture (Kollock [1999]; Raymond [1999]). The term has been borrowed in an attempt to describe the way people contribute FOSS, but originally it denoted those societies in which the economy is neither based on the use of money, nor on barter. Instead, goods are exchanged without the expectancy of any immediate return. Anthropologists, to give a few examples, have observed gift cultures in some native communities of Oceania and even in the Burning Man, a festival that is held yearly in the Nevada desert, in the United States.

The incentive to contribute in this case is the expectancy that your own peers will reciprocate. In fact many FOSS projects start from a personal need of their founders (*“every good work of software starts by scratching a developer’s personal itch”*, notes Raymond [1999]) and are released freely in the hope that others will find them useful, and thus will contribute back.

Gifts aside, reciprocity is a very important mechanism for fostering cooperation within social dilemmas (Axelrod and Hamilton [1981])⁸ Social dilemmas find application in a wide range of real-world situations; we shall cover them properly in the next chapter.

Free Open Source Software can also be viewed as a specific social dilemma, the tragedy of the commons Hardin [1968]; Ostrom [1990], in that software can be thought of as a common good, and individuals can choose to cooperate, i.e. contribute source code and bug reports, or just take a free rider stance and simply draw from the common resource, i.e. use the software.

If contributors are actually doing so expecting other people to reciprocate, then, it would appear that FOSS communities are indeed solving this particular social dilemma. However, there are important differences between the tragedy of the commons and FOSS, which makes it difficult to apply the former to explain the latter (Lerner and Tirole [2002]; Benkler [2002]). The most important one, perhaps, lies in the nature of the common: information is a non-rival good, that is, by its own nature, sharing it does not deplete it. So free-riding is not so much a problem in this case.

Nonetheless, social dilemmas might still be of use to understand how other peer production systems work. For example, peer-to-peer (P2P) file sharing

is even more problematic to establish, this is certainly the case, but in other Peer production experiences, for example programming contests (e.g. Netflix), this kind of incentives are very likely to play an important role.

⁸The latter is a concept from game theory: a social dilemma is a situations in which a rational strategy tells individuals to behave non-cooperatively even though, if all agents – or a majority of them – cooperated, then they would maximize their payoffs.

communities such as Gnutella appeared to show a good deal of free-riding behavior in it (Adar and Huberman [2000]). This was hypothesized to be due to several reasons, both behavioral and technological.

As with academics reviewing their peers' scientific work, participation in FOSS carries, according to economists, an opportunity cost, that comes in the form of the time a developer spends contributing rather than following other pursuits, like his normal job. There are several benefits, both intrinsic and extrinsic, that can counter this cost (Lerner and Tirole [2002]; Benkler [2002]). The first is that contributing to FOSS might help one to better perform in everyday work. Immediate hedonistic incentives are also not so difficult to imagine: for many FOSS contributors contributing to free software is a creative act, and thus rewarding in purely intellectual terms. Indeed writing beautiful code, or a finely composed article about your own favorite philosopher or baseball player can be quite satisfying.

This reward is multiplied when we think that the fruits of this kind of labor are going to be potentially enjoyed by a larger crowd than a circle of friends and acquaintances, which exemplifies how participating can be an extrinsically rewarding activity too.

Other psychological theories have been advanced too, but not always with a clear research agenda. Rafaeli and Ariel [2008] ask whether the concept of self-actualization by Maslow [1954] could apply to peer production contributors. This theory, which still enjoys some popularity among psychologists, is based on the idea that the needs of an individual are structured according to hierarchy that goes from simple physiological requirements (need to sleep, eat, etc.) up to a transcendental urge for self-improvement – termed self-actualization. Rafaeli and Ariel [2008] also look at the uses and gratification people might seek as audience of a media platform. A different approach, based on Activity theory, is taken by Slattery [2009].

But perhaps the most compelling benefits for participating in FOSS projects are in terms of career concerns: firms often hire top contributors of projects they are interested in, in order to let them develop these pieces of software further (Lerner and Tirole [2002]). So there are similarities with peer-reviewing in the scientific community. Besides better job opportunities, access to venture capital is sometimes credited as a form of delayed benefit (Lerner and Tirole [2002]).

Do these findings generalize outside of FOSS? The picture is quite different when it comes to other peer production systems. For Wikipedia participants, for example, surveys show that delayed benefits like improved career opportunities are not as important as other factors (Rafaeli and Ariel [2008]). The general picture is not as definite as with FOSS but the most cited are: the fun of partici-

pating; the sense of belonging to a community; and the desire to fulfill the ideal of Wikipedia of bringing knowledge to the world for free.

Most of these incentives we have seen have a clear sociological nature. Yet, the picture they provide is only meaningful at a macroscopic level – sense of community, gratification from taking part in a thriving project, etc. If we are to investigate, for one, the sense of belonging Wikipedia editors have to their community we need to understand in the first place how such a social group came about, how it formed its identity and its norms. In order to properly understand the phenomenon of peer production, it is thus germane to shift our attention to the microscopic level of the social group. Focusing on the interactions between individual participants will let us understand better what is going on at the macroscopic level of the social group.

2.4.2 Social psychology of user participation

Social psychology is the discipline that studies how the psychological factors of an individual are affected by interaction with a social group. (Sherif [1936]; Festinger [1950]; Festinger and Thibaut [1951]; Lazarsfeld and Merton [1954]; Tajfel [1982]; McPherson et al. [2001]; Friedkin [2001]). In the context of online peer production, social psychology can provide the relevant framework to understand the phenomenon of group formation and evolution (Backstrom et al. [2006, 2007]; Palla et al. [2007]; Ren et al. [2007]). Several researchers looked at motivations for participation that are related to fundamental social aspects of group dynamics. Postmes et al. [2001] looked at the effect of anonymity on group behavior; Song and Kim [2006] the process of acceptance of new members in a virtual group; on the same topic, Ren et al. [2007] investigated the applicability of common identity theory and bond theory to understand the acceptance of new members. The phenomenon of groupthink in online social networks was studied by Hui and Buchegger [2009]. Critical mass theory has been proposed by Raban et al. [2010] to explain the motivation to participate in an Internet Relay Chat (irc) channel. Regarding Wikipedia, Rafaeli and Ariel [2008] propose its mediated interactivity as a possible motivator for participation, while recently Choi et al. [2010] explored the applicability of socialization theory. The growth of wiki communities has been analyzed by Roth et al. [2008]. Wikipedia, the most famous wiki community, was for some time growing at near exponential rates, likely because of its enormous growth of popularity in the period 2006-7 (Suh et al. [2009]). Reagle [2007b] used an ethnographic approach to understand the viability of Wikipedia and argues that the most important characteristic that make it viable project is its collaborative culture, a by-product of

its community norms. Van Alstyne and Brynjolfsson [2005] defined measures of information aggregation and community affiliation to understand the problem of integration within online communities.

Online communities are also appealing because they provide the opportunity to perform experiments. The seminal work of Watts and coworkers on the inequality and unpredictability due to social influence are an example of this methodological feature of online communities (Salganik et al. [2006]; Salganik and Watts [2008, 2009])

Connection with opinion dynamics

As we saw, social psychology provides a rich framework for studying the problem of user participation in peer production communities. For the purpose of the present research, two theories from social psychology, in particular, are relevant: social judgment theory (Sherif [1936, 1961]), and self-categorization theory (Turner [1989]). The first is concerned with understanding under what condition a message will be able to elicit a change of attitude in a subject. A change may be the result of normative reasons, as in our case, or informational reasons, as in the case of a panel of experts trying to reach a consensus on a shared evaluation. A good example of the latter is a jury. As no individual evaluator has perfect information on the verdict, a juror will adjust his evaluation according to the signal coming from his peers. This of course will be a function of the credibility of the other jurors, how extreme their positions are, etc. (cf. Raafat et al. [2009]).

The second is concerned in establishing under what conditions an individual identifies as a member of a social group or not, and how categories (i.e. groups) are formed based on the perceived difference in the relevant cognitive stimuli (Sherif [1961]). Perhaps one of the main tenets of this theory is that there is not a static categorization, but rather that categories depend on the context in which one perceives the relevant stimuli. This idea is embodied in the meta-contrast ratio, also called principle of meta-contrast, which holds that an individual will perceive himself as part of a group based on the ratio between the perceived differences with the other members of the group (ingroup), and the differences between the group members and those outside of it (outgroup). Paraphrasing Salzarulo [2006], two persons from different parts of the Italian-speaking region of Switzerland will not feel similar to each other if they meet on the shores of lake Ceresio in Lugano, in the Italian-speaking Canton Ticino, but they would probably feel close if they met at the central train station of Zürich, in the German-speaking part of Switzerland.

As we saw before, the field of opinion dynamics is specifically concerned with studying phenomena such as group consensus by formalizing in mathematical language concepts such as the meta-contrast principle or the attitude change of social judgment theory (Deffuant et al. [2001]; Hegselmann and Krause [2002]). The concept of bounded confidence is perhaps one of the most elegant and simple formalizations of these ideas.

On a related note, the idea that crowds can lead to an aggregated evaluation or prediction, that it is often better than the one an expert would come up with, is now popular; even though it was known already to early statisticians like Galton [1907], it now goes under the names of “wisdom of the crowds”, “smart mobs”, and “group intelligence” (Surowiecki [2004]).

The wisdom of crowds effect has been invoked, among other things, to explain why Wikipedia, and in general peer production, works. (Tapscott and Williams [2006]) An argument similar to the Central Limit Theorem from probability theory has been, in fact, proposed to explain why many aggregated evaluations given by non-experts could be more precise than a single expert (see references in Lorenz et al. [2011]). However, the conditions of Central Limit Theorem require individual observations to be statistically independent of each other, which is hardly the case, since we are considering interacting agents! Or, in the words of Shalizi [2008]:

Taken seriously, this explanation implies that our economy, our sciences and our polities manage to work despite their social organization, that science (for example) would progress much faster if scientists did not collaborate, did not read each others’ papers, etc. While every scientist feels this way occasionally, it is hard to take seriously. Clearly, there has to be an explanation for the success of social information processing other than averaging uncorrelated guesses, something which can handle, and perhaps even exploit, statistical dependence between decision makers.

In fact, we know that sometimes crowds may not function well: group think may prevail, or over-confidence may develop (Hui and Buchegger [2009]; Lorenz et al. [2011]). The reason why social information processing works so well, whether social influence is beneficial or detrimental to the performance of a group is, thus, still an open, and very important, research question (Shalizi [2008]).

The reason for this excursus is to let the reader appreciate the difference, from a conceptual point of view, between the purported existence of a “wisdom of the crowds” effect and the problem of ascertaining the viability of peer

production: the present research is not interested in the phenomenon of group consensus in a problem-solving context, but as a phenomenon underlying the formation of a community of peers that may, in some cases, engage in forms of problem-solving activities.

2.5 The research on Wikipedia

2.5.1 Wikipedia as a case study on commons-based peer production

Wikipedia has been one of the favorite case studies for researchers interested in peer production for several factors. First and foremost, its open nature has made it possible to anybody who has access to the Internet to collect data on it.

Ciffolilli [2003] was among the first to note its fundamental difference with other communities of practice and highlighted how traditional economics theories about teams and club goods could not apply to it. Lih [2004] proposed it as a tool for participatory journalism, while Voss [2005] presented some early statistics about growth and network structure.

User participation has been extensively studied from an empirical point of view Bryant et al. [2005]; Kittur, Chi, Pendleton, Suh and Mytkowicz [2007]; Kittur, Suh, Pendleton and Chi [2007]; Ortega and Gonzales-Barahona [2007]; Ortega and Izquierdo-Cortazar [2009]; Panciera et al. [2009]; Yang et al. [2010]. Information visualization techniques have been especially useful in understanding how the collaboration patterns unfolded. Viegas, Wattenberg and Dave [2004]; Viegas, Wattenberg and Mckee [2004]; Viegas et al. [2007]; Spinellis and Louridas [2008].

Two main areas of research have emerged over the years, when it comes to discussing Wikipedia: how it governs itself, and how we can make sure that content is accurate. As Wikipedia entered its 11th year of operations early this year, authors are increasingly asking whether the openness of the project is always going to be sustainable or, as Goldman [2009] argues, whether Wikipedia will have to sacrifice its original philosophy in order to survive.

2.5.2 Governance and work organization

As Benkler [2002] points out, an appealing feature of commons-based peer production over firm- or market-based solutions is the capability to solve the problem of task allocation. Under the right conditions about the nature of the tasks

to perform, and whenever a project is able to attract enough workforce, any task will find the best person and the best set of resources to solve it.⁹ In this sense, peer production is the evolution of the concept of communities of practice. These are “*groups of people informally bound together by shared expertise and passion for a joint enterprise*” (cf. Wenger and Snyder [2000]) who form within an organization or an institution and sometimes are able to tackle those transversal problems that cannot be addressed by a single department or branch.

In contrast, peer production communities need not live anymore within the boundaries of a single organizational, and therefore have been praised as a better version of the communities of practice (Huberman and Hogg [1995]; Hogg and Huberman [2008]).

However, large peer production communities have seen the resurgence of governance problems. In the case of Wikipedia, for example, Kittur, Suh, Pendleton and Chi [2007] reported that editors were increasingly being involved in maintenance tasks, and Beschastnikh et al. [2008] found that the usage of policies has increased through the years. Because of its intrinsically bottom-up structure, several other aspects of the governance of the Wikipedia project have been investigated (Burke and Kraut [2008]; Cosley et al. [2005]; Forte and Bruckman [2008]; Hu et al. [2009]; Kittur et al. [2009]; Kriplean et al. [2007]; Lam and Riedl [2009]; Lam et al. [2010]; Leskovec et al. [2010]; Rainie and Purcell [2011]; Reagle [2007a]; Végas, Wattenberg and Mckee [2004]).

As Reagle [2007a] points out, the daily operations that run Wikipedia work by consensus through informal discussion of self-appointed teams. The process for deleting unwanted articles is an example of this (Lam and Riedl [2009]; Taraborelli and Ciampaglia [2010]; Lam et al. [2010]), or the process for promoting users to administrative status (Burke and Kraut [2008]; Leskovec et al. [2010]). These processes are thus open to all the classic problems related to judgment and decision making, such as herding groupthink, aggregation etc. (Plous [1993]; Ottaviani [2001]; Raafat et al. [2009], which affect traditional decision boards such as juries and committees of experts Gilliland et al. [1998]; Daughety and Reinganum [1999]; Visser and Swank [2007].

⁹Both Benkler [2002] and Tapscott and Williams [2006] stress that a problem is a good candidate for solution by peer production methods only if easily decomposable into smaller tasks, and if it is to integrate back the solutions to said tasks. Benkler [2002] also adds that it is important that the granularity of this decomposition is heterogeneous, so that people can find tasks suited for their skills – so that the brightest do not get fed up easily. Whether these post-hoc prescriptions are actually important has never been tested experimentally; on the other hand, the problem has never been raised in the literature.

2.5.3 Quality and the epistemic problems of peer production

As we noted in the introduction, peer production communities challenge the traditional notion of expertise. Assessing the quality of the contributed contents is therefore a crucial problem. Wikipedia is a perfect example of this. As the controversy between Nature (cf. Giles [2005]) and the Encyclopædia Britannica [2006] testifies, the problem of the accuracy of contributions to Wikipedia is far from being solved. According to Denning et al. [2005], information from Wikipedia should be handled with extreme care. Its huge popularity among the general public (cf. Zickuhr and Rainie [2011]), however, seems to indicate the opposite: people consult Wikipedia as any other digital information source available online.

It does not come as a surprise, then, that the scientific community has devoted a lot of attention to the problem of information quality in Wikipedia. Lih [2004] analyzed it as a journalism tool, Wilkinson and Huberman [2007] find a correlation between the level of cooperation and the quality of pages. Priedhorsky et al. [2007] analyze the life cycle of contributed content, from insertion to deletion. Druck et al. [2008] use machine learning to predict the quality of individual contributions. Suh et al. [2008] propose a dashboard tool to improve the accountability of contributors. Stvilia et al. [2008] analyze the collaboration structure from the point of view of content quality. Potthast et al. [2008] also use machine learning, but for vandalism detection. Anthony et al. [2009] assess the connection between user reputation (registered versus anonymous), and reliability of contributions, with surprising results. Hu et al. [2009] try to predict the nomination of pages to the front page of the website. Wöhner and Peters [2009] propose various article life cycle metrics. Liu and Ram [2010] classify contributors based on their role in the collaboration process. Javanmardi et al. [2010] model user reputation similarly to Adler and Alfaro [2007], in order to classify edits made by users with a high standing.

All this research is eminently empirical, and in fact, in the absence of clear definition of content quality, many of the above studies set out to measure proxies of it. From a more theoretical point of view, the problem of quality of information is a problem of social epistemology, and as such should be considered (Sanger [2009]). In this context it is worth mentioning the work of Roth and Bourguine [2005] on epistemic communities and the simulation approach of Hegselmann and Krause [2006] on the cognitive division of labor. As more and more research is devoted to the problem of quality, the need for epistemic models of content creation is surely needed.

Chapter 3

Empirical analysis of user activity lifespan

It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

Sherlock Holmes

3.1 Introduction

In this chapter we begin our analysis of user participation, focusing on the case study of Wikipedia, the free online encyclopedia. The simplest quantity of user participation we can study is the period during which a user contributes to the digital common. We call this quantity the activity lifespan of a user, or user activity lifespan. How can we measure the user activity lifespan in Wikipedia? There is both good and bad news. The good news is that we can easily collect data about the contributions of each registered user of Wikipedia. Snapshots of the Wikipedia database are routinely dumped to file and released to the public by the Wikimedia Foundation, who manages the infrastructure on which Wikipedia runs, alongside with its sister projects. This means that we can collect from the actual logs of the Mediawiki software the activity period associated to each registered user account.

The bad news is that this quantity is not precisely the user lifespan. Since Wikipedia allows anyone to edit its content, many contributions are made by

anonymous editors. There is large anecdotal evidence that an initial period of anonymity is common among Wikipedia editors.

There are a number of reasons for this. Several people, for example, come into contact with the project by contributing anonymously, and only after some time do they notice the possibility of registering a user account. Others prefer to stay anonymous at first, for example to gain confidence with the website. Other contributors prefer to remain anonymous for their full period of participation. Aside from the one-time contributors, in all other cases this behavior is called “lurking” (a use of the term that originates in early Internet forums). There is very little we can do about this situation, so we’ll have to keep in mind this inherent bias when drawing our conclusions about user participation.

The second problem with inferring user participation from activity logs is that these data might not be accurate. For example, a user might forget to log into the wiki software, so we might miss some of his contributions. Since we are actually measuring the whole lifespan of user activity, missing a few contributions in the middle of the activity period is not a real problem, but we have to be careful about the tail end. In principle, this problem would be a real concern only for a class of users – those whose activity lifespan is relatively short compared with the real core members of the community. However, since the distribution of inactivity periods (i.e. the inter-edit time distribution) decays with a heavy tail (Radicchi [2009]), we cannot exclude inaccuracies also for other kinds of users. To curb this problem, we will further restrict our attention to users who are inactive. How we define inactivity, and how this affects the results of our statistical analysis, will be explained in detail later.

3.1.1 Empirical findings

Multimodal distribution of user activity lifespan

In all cases, we found the data to be compatible with the superposition of two or more truncated log-normal distributions. A power-law distribution is instead rejected by the data. An interpretation for this phenomenon is that at least two different regimes govern the participation of individuals to these versions of the Wikipedia project: occasional users, who fail to maintain interest in the project after the first few attempts to contribute, and expert users, whose deactivation is probably more related to external factors like the loss of personal incentives in contributing and similar. While for the former the history of editing is roughly a time scale of hours, the latter have a record of participation that is rather measured on a scale of years.

Evolution of user participation

Using our model, we characterize how the participation of users over time evolves, as the system ages up. We find that the statistical description of the one-timers is stable over time, while the properties of the group of expert users change as a consequence of the aging of the system (see figures 3.7–3.9)

Maximum inactivity

We find evidence that the inter-edit time distribution decays with a heavy tail. Since our analysis relies on preprocessing the dataset in order to select only “inactive” users, we check that the results of our analysis is not affected by the choice of the parameter used for determining when a user is inactive; we find that for the one-timers it has no quantitative effect. For long-term users the statistics changes smoothly with the parameter. Thus the quantitative results here depend on the arbitrary choice of the parameter, but there are not qualitative changes once a value of the parameter large enough is chosen.

3.2 The dataset

3.2.1 Data collection

Wikipedia comes in different languages, and most existing languages have their own Wikipedia, or localized version.¹ However, these versions are not mere translations: each language version of Wikipedia has a community on its own, so we are actually dealing with many, largely independent realizations of the same peer production community. We choose to focus our attention on five of the largest Wikipedia communities. These are: the English (at the time of writing this dissertation, it has 3,777,000+ articles), German (1,340,000+), French (1,165,000+), Italian (853,000), and Portuguese (702,000+).

Our data come from the official database dumps released by the Wikimedia Foundation.² These dumps contain snapshots of selected tables from the databases of all localized projects of Wikipedia. Each dump comprises all data from previous dumps of that locale. Dumps are produced periodically, although

¹A “locale” is, in technical jargon, a specific language or language group. At the time of writing there are 282 localized versions of Wikipedia. See https://meta.wikimedia.org/w/index.php?title=List_of_Wikipedias&oldid=3029166.

²See <http://download.wikimedia.org>.

rev_user	rev_user_text	rev_timestamp
7077	Moroboshi	2006-05-26 04:37:45
0	82.50.4.229	2006-05-26 19:15:38
36426	Sailko	2006-06-05 08:32:48
57872	Dapa19	2006-06-07 16:31:58
35813	Moloch981	2006-06-07 20:24:14

Table 3.1. Data excerpt from the database of the Italian Wikipedia. See the text for more information.

the frequency with which a dump is produced depends on its size. The dumps we analyzed are from 2009.

Since dumps are released publicly, and not for mere research purposes, they contain only the kind of public information that the general public would be able to access through the Web interface of Wikipedia. Any piece of personal information about users – which is stored in a specific table called “user” – is not included in database snapshots. Thus we do not have access to the registration time of user accounts, nor to the time of last login.³ We do have access, however, to the history of page contributions of each user, which is stored in the table “revision”, and from this we can infer the lifespan of user participation.

Below we report the values of the two variables together with the user name or, in the case of anonymous contribution (“rev_user” = 0), the IP address of the contributor (column “rev_user_text”). The data refer to ten consecutive revisions from the Italian Wikipedia made to the page about Pope Clement VII (*1478, †1534), in the period June–July 2006.⁴ The data we use for our analysis is thus the history of user contributions to pages. The database records some metadata on each page revision. Of these, we use only two variables: the numeric ID of the contributing user (column “rev_user”), and the time stamp of the revision (column “rev_timestamp”).

Our raw data is thus a sequence of revisions (u, t) where $u > 0$ is an integer that represents the u -th user and t is expressed in seconds. We can consider the individual user time series $t_1^u < \dots < t_{N_u}^u$, where N_u is the total number of edits of user u . The time interval of observation of our data is thus $[t_0, t_1]$ where

³Mediawiki, the wiki software on which Wikipedia runs, does not keep track of this information, but just of the last time made any change to the site, including logins and normal contributions. Unfortunately, those users who authenticate themselves via cookies are not tracked by this mechanism.

⁴The contributions show in table 3.1 can be seen here: https://it.wikipedia.org/w/index.php?title=Papa_Clemente_VII&action=historysubmit&diff=3283064&oldid=3275026.

$$t_0 \equiv \min_{u>0} \{t_1^u\} \text{ and } t_1 \equiv \max_{u>0} \{t_{N_u}^u\}.$$

3.2.2 Statistical considerations

The data collection is affected by the two following statistical biases:

1. Selection bias: all anonymous contributions, in particular, are mapped to the same user ID, and thus we have to discard them. Our results therefore apply only to registered users and not to anonymous users.
2. Sampling bias: because of the way we compute the duration of user participation, we have to restrict ourselves to users that performed at least two edits. This may under-represent short-time user participation, as users that perform only one edit are also likely to stop participation soon after that.

Moreover, we should also take into account the two following characteristics of our data:

1. Interval censoring: since we observe participation durations up to time t_1 , our data are affected by right-censoring, that is, the participation duration of certain users may actually be longer than the one we get by simply computing the time span between their first and last edits.
2. Data truncation: our data are also left truncated for obvious physical reasons: a human editor cannot perform two consecutive edits at arbitrarily high speed.

Since handling a dataset that is both right-censored and left-truncated can be complicated, we decided to left-truncate our data and work with a fully truncated sample. The left-truncation should keep only observations that are not affected by the left-censoring phenomenon, that is, those users that are fully inactive. Of course, we have no guarantee that a user who has stopped editing will not start again at any time in the future. So we have to perform some kind of data preprocessing. Specifically, we seek a method to filter those users that are likely to be still active outside of the observation interval. To do this, we define a maximum inactivity time span τ_{inact} and say that a user is definitively inactive if the time elapsed since his (or her) last contribution $t_l^u \left(\equiv t_{N_u}^u \right)$ is more than τ_{inact} , that is, any user u such that $t_1 - t_l^u > \tau_{\text{inact}}$.

The maximum inactivity time τ_{inact} should be not too small, otherwise it might miss many active users, and of course not too long, or it will wrongly classify as active users that are not.

3.2.3 Robot users

On the Web robots (“bots”, in jargon) are pieces of software that interact with a website through a programmatic interface, or simply by sending out HTTP commands. In the case of Wikipedia, bots are given a specific user account, and are often used to perform repetitive tasks, such as fixing templates, checking hyperlinks, and so on.

Because of the massive number of edits that certain bots are capable of performing, these accounts are usually filtered out when analyzing the distribution of edits of users or similar quantities. Lists of known bots are available online, and the Mediawiki database has a table (“user_groups”) for recording special users that contains information on all robot users.

In our case, bots do not constitute a big problem for our statistical analysis. First, the lifespan of user participation has little to do with the total number of edits. Second, each Wikipedia contains only a handful of bots, so even including them will not affect noticeably the statistics on user lifespan.

3.3 The Model

Here we review briefly the theory behind truncated Gaussian mixture models (TGMM). A truncated distribution is obtained by taking a continuous distribution and restricting its support to a connected subset of its support. In the case of univariate distributions the original support is usually \mathbb{R} and the subset is an interval with extremes $a < b$, where $a, b \in \mathbb{R} \cup \{\pm\infty\}$. If one of a or b is not finite, one speaks of a truncated distribution on the *right* respectively *left*. A truncated distribution is thus specified by considering the original distribution conditional on being in (a, b) ; for the normal distribution with location μ and scale σ the probability density of its truncated version is thus defined as:

$$p(x) = \begin{cases} \frac{[\Phi(b') - \Phi(a')]}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) & \text{if } x \in (a, b) \\ 0 & \text{else} \end{cases} \quad (3.1)$$

where Φ denotes the cumulative distribution function of the standard Gaussian

$$\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt = \frac{1}{2} + \left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \quad (3.2)$$

and the extremes are standardized, i.e.

$$a' = \frac{a - \mu}{\sigma} \quad (3.3)$$

and similarly for b' . Expressions for the first and second centered moments of the truncated normal exists, but rely on the knowledge of the sufficient statistic of the underlying Gaussian (cf. Johnson et al. [1994]):

$$E[X|a < X < b] = \mu - \sigma \frac{\phi(b') - \phi(a')}{\Phi(a') - \Phi(b')} \quad (3.4)$$

$$V[X|a < X < b] = \sigma^2 \left[1 - \frac{b'\phi(b') - a'\phi(a')}{\Phi(b') - \Phi(a')} - \left(\frac{\phi(a') - \phi(b')}{\Phi(b') - \Phi(a')} \right)^2 \right] \quad (3.5)$$

here ϕ refers to the density of the standard Gaussian distribution.

A mixture model is the superposition of two or more distributions, so that the resulting density is a weighted average from several components:

$$p(x) = \sum_{k=1}^K \pi_k p_k(x; \theta_k) \quad (3.6)$$

here $p_k(x; \theta_k)$ is the density of the k -th component, with vectors of parameters θ_k , evaluated at x . If we denote with π the vector of weights and call $\theta = (\theta_1 \dots \theta_K)$, then the complete set of parameters of the mixture is given by (θ, π) . In order to consistently estimate the parameters of this model, one can follow a maximum-likelihood approach and set oneself to maximize the log-likelihood $\mathcal{L}_x = \log \prod_{i=1}^N p(x_i)$, so that $\hat{\theta} = \arg \max_{\theta} \mathcal{L}_x$ is the MLE estimator of the parameters of the mixture. This is, however, often infeasible using constrained maximization methods because of the peculiar behavior of the log-likelihood function (cf. Bishop [2006]). A solution to this problem comes in the form of the expectation-maximization algorithm (EM). In fact, the assumption that the density is given by equation 3.6 is equivalent to assuming a generative model of the data that uses the information of a set of latent variables $\mathbf{z} = \mathbf{z}_1, \dots, \mathbf{z}_N$. The latent variable \mathbf{z}_i has a 1-of- K representation:

$$z_{ij} \in \{0, 1\} \quad j = 1 \dots K, \quad \sum_j z_{ij} = 1 \quad (3.7)$$

and is used to determine the index of the components that was responsible for generating the value of the i -th observation (cf. Dempster et al. [1977]). Once

this representation is used, it is possible to simplify the log-likelihood function for the full data (i.e. latent variables included); in particular, when using densities from the exponential family of distributions for the $p_k(x; \theta_k)$, the logarithm can be simplified out, a thing that otherwise would not be possible due to the presence of the weighted average in equation 3.6.

Of course, in practice, one does not observe the latent variables \mathbf{z} and here's where the key intuition of EM lies: it is possible instead to consider the expected log-likelihood $E_{\mathbf{z}}[\mathcal{L}_{\mathbf{x}, \mathbf{z}}]$ (E-step) and maximize it separately, thus finding MLE estimates of the parameters of the model (M-step). The expectation is taken with respect to the distribution of the latent variables $P(\mathbf{z}|\mathbf{x}, \theta^{(s)}, \pi^{(s)})$, where s is the current step of the algorithm.

This is easily computable, using Bayes theorem, from the current assignment of observations to components.⁵ New estimates $\theta^{(s+1)}, \pi^{(s+1)}$ computed at the M-step thus refer to the current previsions of the latent variable at the s -th iteration of the algorithm. This iterative procedure is guaranteed to converge to a local maximum of the log-likelihood function.

In our case, however, we also have to take into account the truncation from equation 3.1 when computing the estimate of the parameters for each component in the M-step. To do this, we do employ an approximation. We first compute the weighted estimates of $\hat{\mu}$ and $\hat{\sigma}$ of the EM algorithm, and then plug the estimate of the variance in the right hand side of equation 3.5 and finally substitute this corrected estimate of the truncated variance, together with $\hat{\mu}$, in equation 3.4.⁶

We find that, in non-pathological cases, this approximation produces asymptotically unbiased estimators. In figure 3.1 we plot how the distribution of residuals of the EM estimator behaves, as the size of the sample used to estimate the parameters grows. We can see that the approximation we used does not affect the quality of the MLE estimator produced by EM. In the case of the weights, because of the constraint that $\sum_k \pi_k = 1$, we only plot the residuals for one of the two weights. The 'troublesome' cases arise when multiple components overlap significantly, so to make them indistinguishable from a single component, or when one or both the truncation extremes fall very shortly (i.e. less than one standard deviation) from one or more components' center. We do not think this

⁵The initial assignment is performed with the k -means clustering algorithm.

⁶Another approach for doing inference for a mixture of truncated distributions with EM is to introduce a new set of latent variables that refer to the full data sample, i.e. together with the observations that are missing due to the truncation (cf. McLachlan and Jones [1988]). This approach, however, takes also into account the grouping of data (instead of the observed values themselves, one gets histogram-like frequencies), which is not our case.

problem endangers the quality of our results too much, as our data do not seem to fall in any of these cases (see table 3.3).

3.4 Results

We downloaded and processed the data in order to produce, for each Wikipedia community, lifespan observations $\tau_1, \tau_2, \dots, \tau_N$. As we said above, we removed robots and anonymous accounts. Finally we filtered out users according to our inactivity criterion. Following Wilkinson [2008], we consider as inactive all users whose last revision dates back more than $\tau_{\text{inact}} = 180$ days from the date of the most recent revision recorded in the data. As we noted above, this does not guarantee that a user classified as inactive will not contribute anymore in the future. Moreover, the choice of τ_{inact} may seem somewhat arbitrary. We performed our analysis also with other choices of τ_{inact} . We found that nonetheless they all lead to very similar results, making this arbitrariness less of a problem for our analysis. We discuss these results later in this section.

Wikipedia	α	τ_{\min}	$n_{>\tau_{\min}}$	p -value
Italian	3.99 ± 0.14	688.53	461	0.09
German	4.84 ± 0.10	1013.89	1342	0.1
French	3.58 ± 0.14	681.01	351	0.09
Portuguese	3.91 ± 0.11	619.37	693	0.08
English	6.95 ± 0.08	1119.43	5376	0.04

Table 3.2. Power law fit. p -values for statistically significant estimates (≥ 0.1) are quoted in bold.

3.4.1 Multi-modality of activity lifespan distribution

Table 3.2 shows the result of the fit of user activity lifespan τ to a power-law distributional model. The exponent α and the starting point τ_{\min} of the power-law distribution are found via MLE, (Clauset et al. [2009]). We use non-parametric hypothesis testing to assess the goodness-of-fit of this model, in particular a Kolmogorov-Smirnov (K-S) test. Other choices like Anderson-Darling could be applied as well, but we decided to use K-S mainly for its intuitiveness.

The p -value from a K-S test tells that the data do not support a power-law model – with the exception of the German, for which there is a weak support.

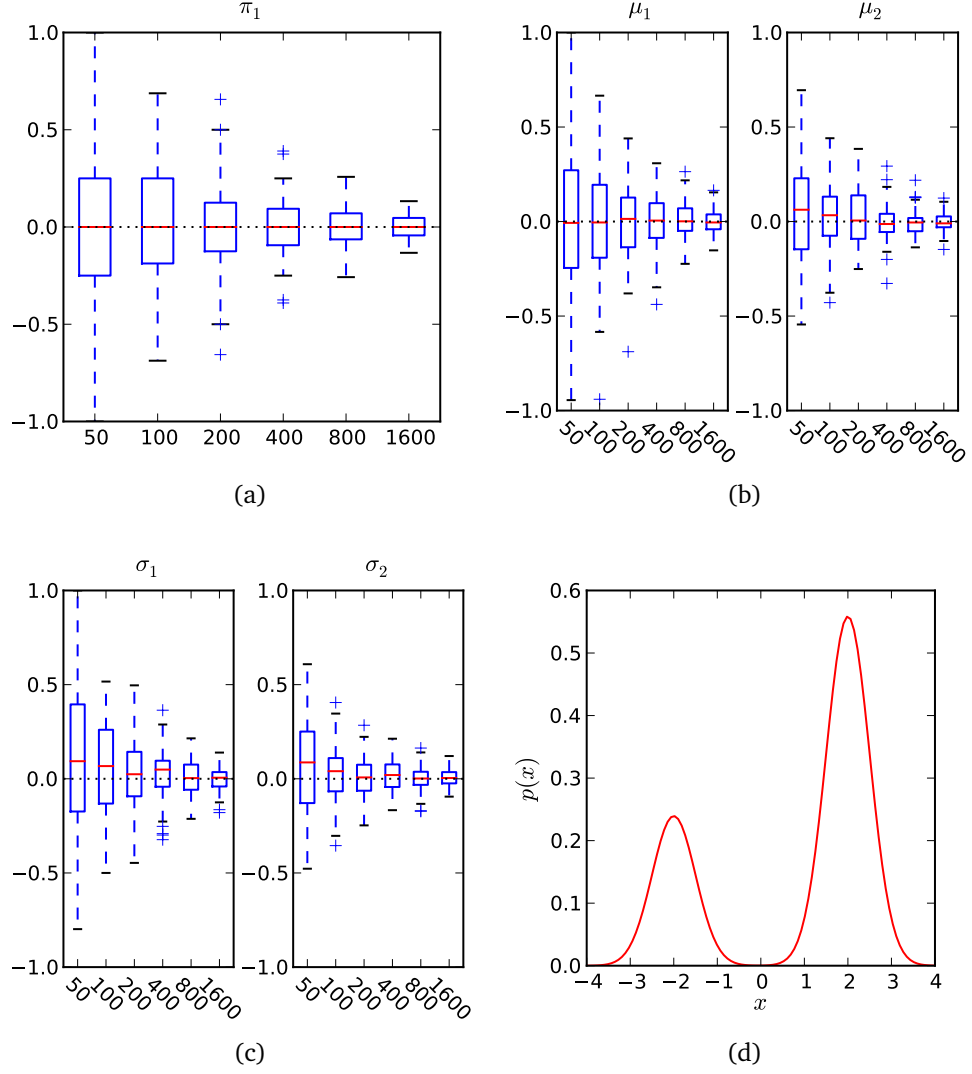


Figure 3.1. Validation of the EM technique. (a)–(c) box and whiskers plot of scaled residuals versus sample size. The box size corresponds to the 1st and 3rd quartile, the red segment is the median and the whiskers are 1.5 times the inter-quantile range (IQR). (d) the model used to compute the residuals, with parameters $(\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1, \sigma_2, a, b) = (.3, .7, . - 2, 2, .5, .5, -4, 4)$.

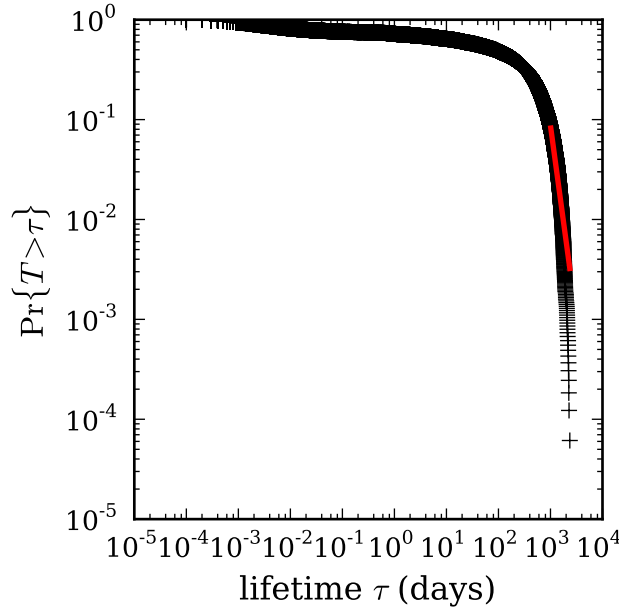


Figure 3.2. Power law fit for the German Wikipedia. Black crosses, empirical CDF; solid red line, model fit.

Moreover, the power law decay is found only in the extreme tail of the distribution ($\tau_{\min} > 600$ days in all cases), which means that a power-law model fails completely to characterize the structure of the data for all but the largest values of τ . Figure 3.2 shows, as an example, the empirical complementary cumulative distribution function for the German Wikipedia together with its power-law fit.

The power-law model is not a satisfactory model for two reasons. As we noted already there cannot be any user whose lifespan in the project is longer than the age of the project itself; this explains the sharp cutoff for high values of τ , which is incompatible with a power-law decay that spans more than an order of magnitude.

Second, the overall distribution of user lifespans might be the product of two (or more) regimes of participation: one comprising users whose interest, after a few edits, fades away quickly. The other one might comprise the so-called expert users, whose motivations for participating are not affected anymore by the daily outcomes of their editing actions, but rather might be tied to some stronger form of incentives (e.g. the ideology of free and open projects). The presence of multiple characteristic scale is incompatible with a scale-free decay.

From a statistical perspective, since the empirical distribution is clearly heavy-tailed, we thus want to test a heavy-tailed distribution that accounts for multiple

characteristic scales. A mixture of (truncated) log-normal distributions, for example, could work.

Since the logarithm of a log-normally distributed variable is itself normally distributed, we can infer the parameters of our truncated mixture model by applying our custom EM technique to $u = \log(\tau)$. Table 3.3 and Figure 3.3 display the results of the parameter estimation. In all cases, a K-S test does not reject the hypothesis that the data are drawn from the same distribution.

Wikipedia	π_1	π_2	μ_1	μ_2	σ_1	σ_2	p -value	(a, b)	date
Italian	0.32(4)	0.68(4)	-5.4(3)	4.3(3)	1.7(3)	1.9(3)	0.688	(-9.2, 7.5)	2009-09-13
German	0.44(3)	0.56(3)	-2.2(2)	5.6(2)	3.8(2)	1.1 (3)	0.632	(-9.2, 7.8)	2009-05-25
French	0.28(3)	0.72(3)	-5.5(2)	4.6(3)	1.8(2)	1.8(3)	0.464	(-9.2, 7.7)	2009-06-16
Portuguese	0.46(3)	0.54(4)	-5.5(4)	3.5(3)	1.5(4)	2.2(4)	0.612	(-9.2, 7.6)	2009-06-17
English	0.47(5)	0.53(5)	-5.3(5)	3.2(4)	1.6(5)	2.2(5)	0.54	(-11, 7.8)	2008-04-16

Table 3.3. Estimated parameters for the truncated normal model. In parentheses the significant digit of the standard error of the estimator. p -values for statistically significant estimates (≥ 0.1) are quoted in bold.

The results of the fit are particularly interesting as there are striking similarities in the estimated parameters across these different languages – the German Wikipedia being the only notable exception. In particular, we note that μ_1 , the average activity lifespan of the short-term users is, for the French Wikipedia, approximately equal to -5.5 , which means an average period of $\langle \tau_1 \rangle = \exp\left(\frac{1}{2}(-5.5 + (1.8)^2)\right) \approx 29.74$ minutes, while for the expert users $\langle \tau_2 \rangle \approx 502.7$ days.

From a visual inspection of the fit of the German data, we note that here EM produced an estimate that fits well the second component of the long-lived users, at the expense of a little or no agreement for μ_1 and σ_1 . Indeed, EM is only guaranteed to converge to a local maximum of the likelihood function and thus we expect this discrepancy to be due just to a deficiency of our estimation technique.⁷

This leads us to ask: what is the best value for the number of components k ? A mixture model can have an arbitrary number of components, which means that we want to compare the goodness of models who have different number of parameters. More parameters necessarily imply a better fit with the data. This is actually a problem because we can run into over-fitting, i.e. the model learns the noise in the data, and performs poorly at generalizing the underlying structure. We thus performed a model comparison by computing the Akaike information criterion (AIC), which is given by:

⁷in order to increase the chances of hitting a good maximum of the log-likelihood function, our estimation procedure is repeated 25 times for each data set.

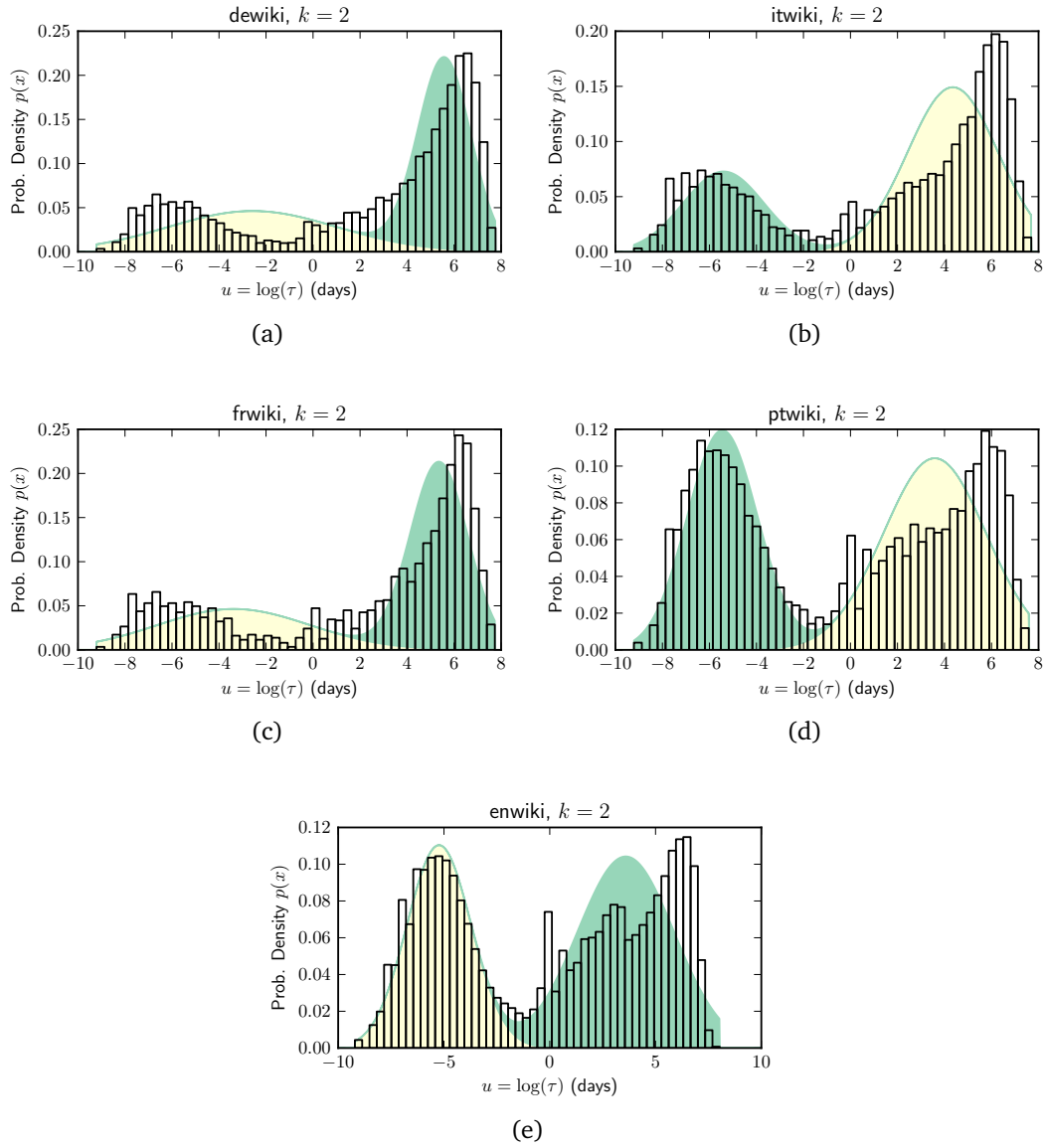


Figure 3.3. Truncated Gaussian mixture model (TGMM) fit. (a) German, (b) Italian, (c) French, (d) Portuguese, (e), English Wikipedia. $\tau_{\text{inact}} = 180$ days. Number of components: 2. Histograms: empirical data. Stacked green/yellow area: TGMM density.

$$\text{AIC} = 2K - 2\log \mathcal{L}^* \quad (3.8)$$

where \mathcal{L}^* is the maximum of the likelihood function and k the actual number

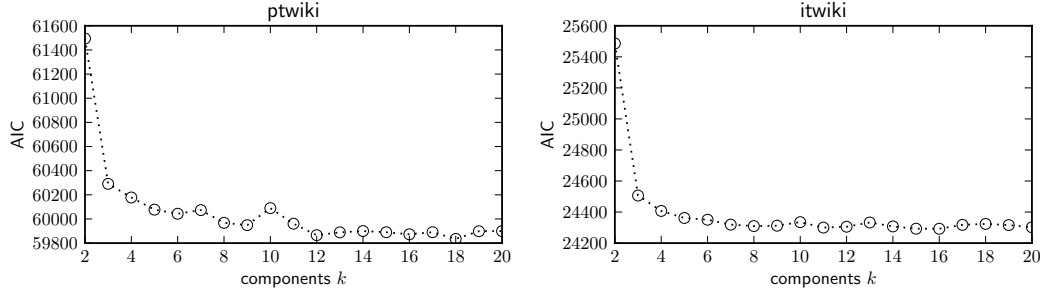


Figure 3.4. Bayesian model comparison. Akaike Information Criterion. Top: Portuguese Wikipedia. Bottom: Italian Wikipedia.

of parameters of the model, which in the case of the TGMM is $K = 3k - 1$. This means that we discount models with a higher number of parameters. The AIC can be interpreted as the loss in information involved in using a given model instead of the true, unknown model that generated the data (under the assumption that such a model exists, of course). This means that we want to favor those models that minimize the discrepancy given in 3.8.

Figure 3.4 shows the results of the model comparison exercise. For simplicity, we report here only the Italian and Portuguese Wikipedia. We can see that the AIC score drops at $k = 3$. For more components, improvements are negligible. So, we can settle for the value at the knee of the curve ($k = 3$).

3.4.2 Inactivity periods

We now turn to the problem of evaluating how much our analysis depends on the choice of the parameter τ_{inact} . A simple binary classification criterion for whether a user i in our sample is inactive is to consider the time from the last revision as test if it above a certain threshold. We classify as inactive those user i for which $t_1 - t_{N_i}^{(i)} > \tau_{\text{inact}}$, where τ_{inact} is the period of inactivity past which the user is likely not to contribute anymore in the future.

Of course this binary classification would be robust under the assumption that individuals contribute to Wikipedia with a characteristic rate of activity throughout their whole life cycle. Once a drop in the activity is observed, there is a high likelihood that the user has turned permanently inactive. The validity of this assumption will be explored in detail in chapter 7. For now, we can ask ourselves whether it is likely or not to see, in the history of contribution of any editor in our sample, any inactive period spanning more than τ_{inact} days.

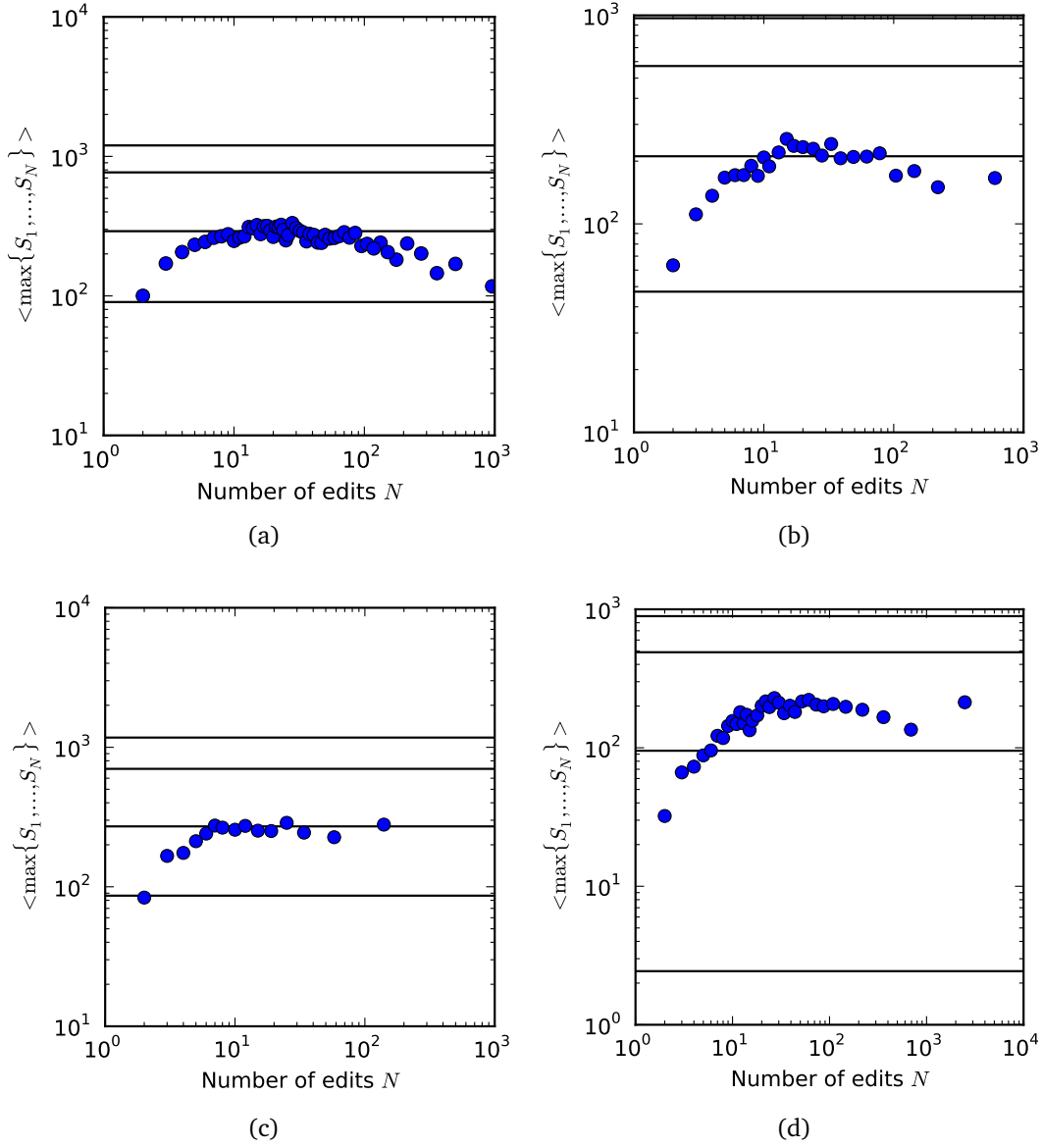


Figure 3.5. Analysis of the inter-edit time distribution. (a) German, (b) Italian, (c) French, (d) Portuguese Wikipedia. Filled blue circles: scaling of the average maximum statistics as a function of the number of edits (i.e. sample size). Axes are on the log-log scale. The grouping is count dependent so that in each bin there are at least 100 observations. The relative standard error bars are all smaller than the data points. Horizontal solid lines: from bottom to top, median, 75th, 95th and 99th percentile of the full distribution of the maxima, i.e. not grouped by number of edits.

Let us consider a generic user who has performed N edits. We denote with S_0, S_1, \dots, S_{N-1} the time intervals between consecutive edits, that is $S_i = t_{j+1}^{(i)} - t_j^{(i)}$. We consider the distribution of $\max\{S_0, \dots, S_{N-1}\}$. Figure 3.5 shows how the maximum grows as a function of the number of edits N . For each value of N the average maximum inter-edit time is depicted. In order to control the fluctuations for high values of N , users are binned by number of edits, with multiple bins merged together so that there are at least 100 observations in each bin.

If the underlying distribution of the inter-edit times had a tail decaying faster than a power-law (i.e. exponentially or like a Gaussian), the maximum would be a slowly increasing function of N (cf. Sornette [2004]). Here, instead, we see that for small values of N the maximum is growing faster than that. For instance, the maximum inactive period of editors with more than 10 edits can be, on average, 6 months long. This is another confirmation that the inter-edit interval distribution of contributions on Wikipedia must be decaying with a heavy tail. For more than 20 edits the maximum levels off, which might be due to some form of exponential cutoff of the tail of the interval distribution.

Therefore, we have to check whether the choice of τ_{inact} affects the results of our estimation. Of course, we know already that it will: each choice of this parameter produces a different dataset, so we expect to see quantitative differences in the results, if we use two different values of τ_{inact} . What we ask ourselves here instead, is whether there are also qualitative differences, for example such that the current number of components is not enough to fit the data accurately and more components are instead needed.

A good way to test this is to see how the sufficient statistic of the truncated Gaussian mixture model changes as the τ_{inact} takes different values, i.e. are the parameters of the mixture very sensitive with respect to changes of τ_{inact} ? This will also allow us to assess what the minimum value of the parameter that is safe to use is. Thus we repeated our analysis of the distribution of the lifespan τ for several values of τ_{inact} . In figure 3.6 We plot the parameters of a truncated mixture as a function of τ_{inact} . For simplicity, we used a model with only $K = 2$ components. As a comparison, we also plot the average lifespan $\bar{\tau}$.

Looking at the effect of τ_{inact} on the mixture means $\mu_{1,2}$ (left y -axis) we can see that the estimates jump in the cases of the French and German. For the German, in particular, this happens approximately at 300 days. This is consistent with the result of the fit of the 2-components TGMM model (figure 3.3a). For the French Wikipedia, the jump happens earlier. For the other Wikipedia versions, it seems instead that τ_{inact} does not elicit any sudden change. This means that,

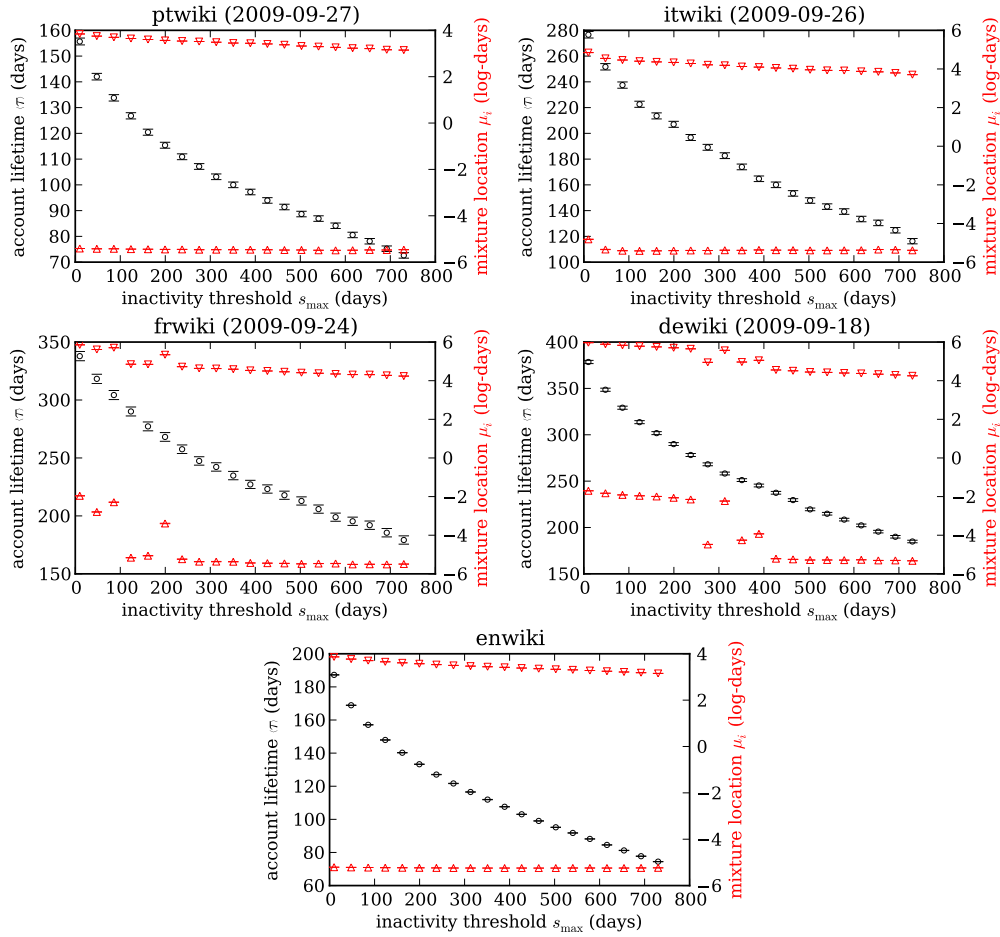


Figure 3.6. Sensitivity analysis of data right-truncation. Estimated mixture means for a truncated Gaussian mixture model with $K = 2$ components (short-term and long-term) as a function of τ_{inact} . Black circles: average user lifespan. Red upward triangles: estimated means of short-term component. Red downward triangles: estimated means of long-term component.

even though the quantitative results of our analysis depend on the actual choice of this parameter, choosing any value larger than (roughly) 300 days ensures that there is no real qualitative change in the structure of our data.

3.4.3 Temporal evolution

As previously remarked, the choice of discarding “active” users by means of binary classification was introduced to transform our right-censored sample into a right-truncated one, thus allowing us only to take into account truncation in the estimation of the mixture parameters. In fact, our data come from a specific observation window, and thus the results of the estimation depend on its size. Thus, for example, setting a smaller or large observation window will change the results of our analysis. But how? We can try to get an idea of whether the distribution of lifespan has reached a stable state by depicting its evolution in the past. More specifically, we can restrict our observation window to any point in the past and compute again estimates of the parameters of the truncated Gaussian mixture model for this reduced dataset. By plotting the evolution of the fitted parameters in time we can see whether they are reached a value regardless of the size of the window or not.

Figures 3.7, 3.8, and 3.9 depict a comparison of the results of this analysis for several choices of the value of τ_{inact} ; in particular, we tested the values of 7, 30, 90, 180, 365 and 730 days. We took yearly-spaced points in the period that goes from an initial timestamp to the last recorded timestamp in our datasets. The initial timestamp is chosen so that the system contained at least 100 users at that time. At each of these dates, we restrict our datasets only to those users whose last revision recorded in the full dataset precedes the date.

The mean of the short-term lifespan component μ_1 does stabilize in all five cases after a few years. That this parameter eventually stabilizes is expected, since the observation window is several orders of magnitude larger than the average short-term lifespan, and thus censoring does not have any effect on these scales. However, it is also interesting to note that the stabilization happens with negative trend, i.e. in early years the short-term lifespan was somehow longer than it is now.

In contrast, the mean of the long-term component μ_2 shows a steady growth. This is basically a consequence of the data censoring with a growing observation window. There does not seem to be any plateau effect, even though we can clearly discern some discrepancies from a pure linear trend, which would be imputable only to censoring.

The evolution of mixture weights $\pi_{1,2}$ gives instead some insight into the

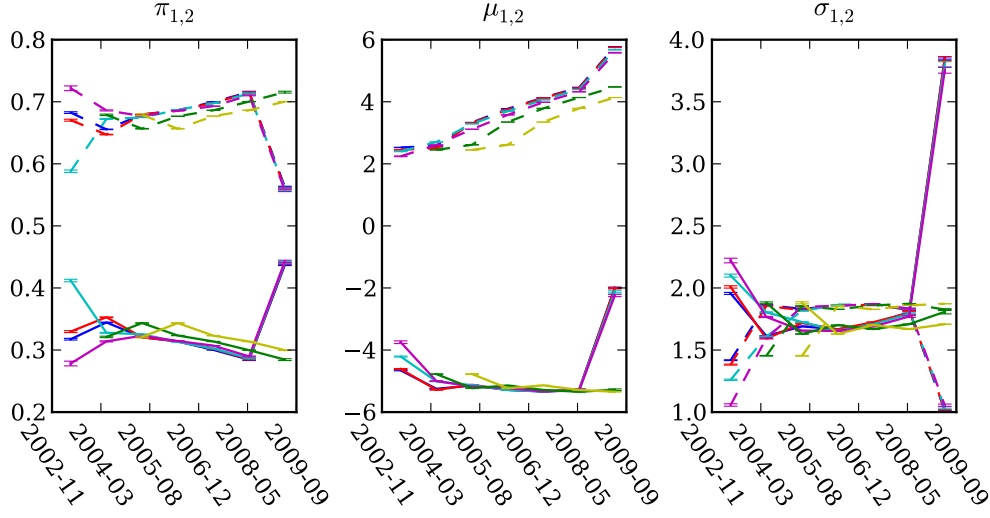


Figure 3.7. Evolution of estimated parameters. German Wikipedia. Colors correspond to different values of the time τ_{inact} used to classify a user as inactive: 7 (blue), 30 (red), 90 (cyan), 180 (magenta), 365 (green), 730 (yellow). Solid lines: π_1, μ_1, σ_1 (short-lived users). Dashed lines: π_2, μ_2, σ_2 (long-lived users). Error bars are the 95% confidence intervals of the estimator.

percentage of users types. It is interesting to note that not all Wikipedia communities have the same values. While the German, French, and, to some extent, the Italian seem to have a 70/30 composition, the English and the Portuguese seem to have a more even balance between short-term and long-term users.

3.5 Discussion

Our analysis gives us a very precious insight into user participation, namely that users belong to different classes and that these differences can be detected quantitatively. In light of these results, it is legitimate to ask what makes a user stay for a short or long period. While the number of new users that join a community at any given time can be largely attributed to the popularity that community is enjoying, once a person has joined, there are still several factors that will influence his or her decision to become part of the community or not. Of course we can immediately count several “exogenous” factors (e.g. how the UI is, how fun it is to perform the tasks). But there could be also other, perhaps endogenous factors, at work. In particular, cultural and social factors are likely

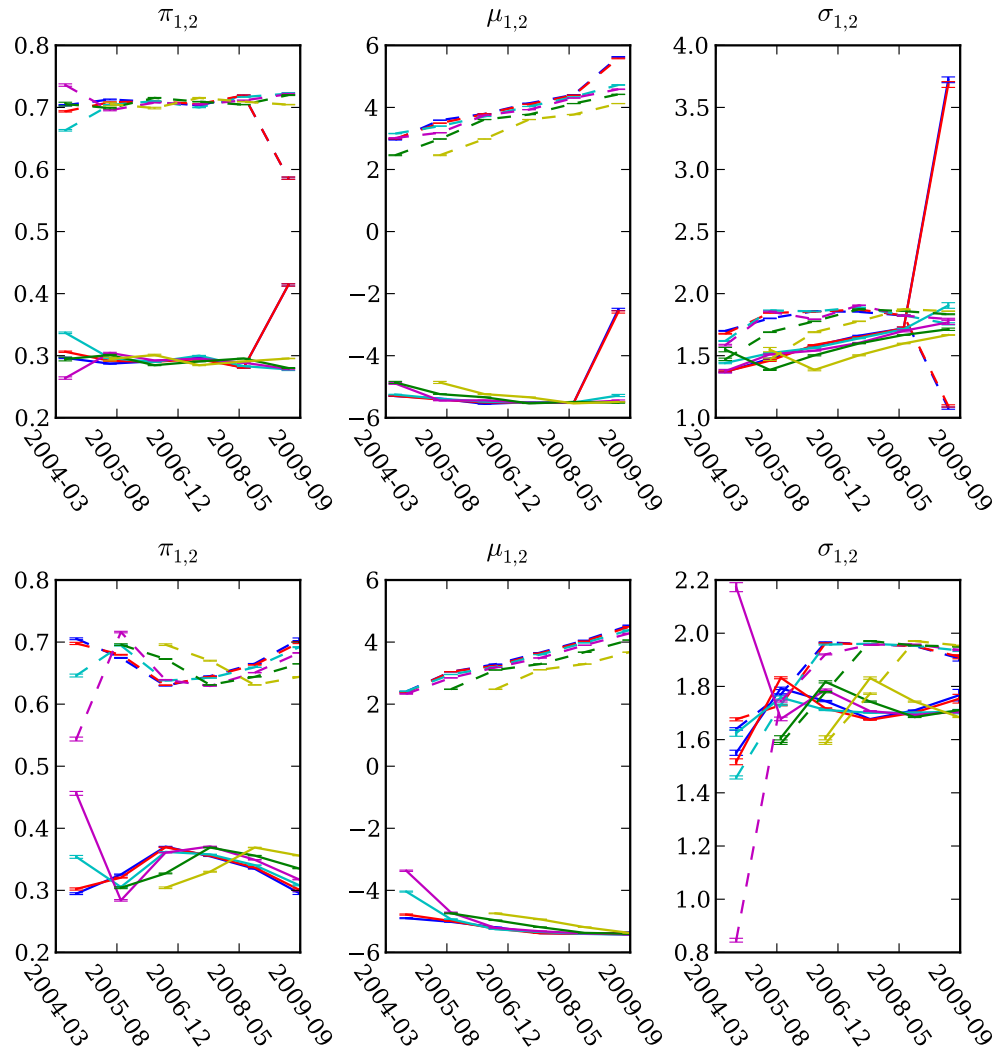


Figure 3.8. Evolution of estimated parameters. Top: French Wikipedia. Bottom: Italian Wikipedia. See 3.7 for explanation of symbols.

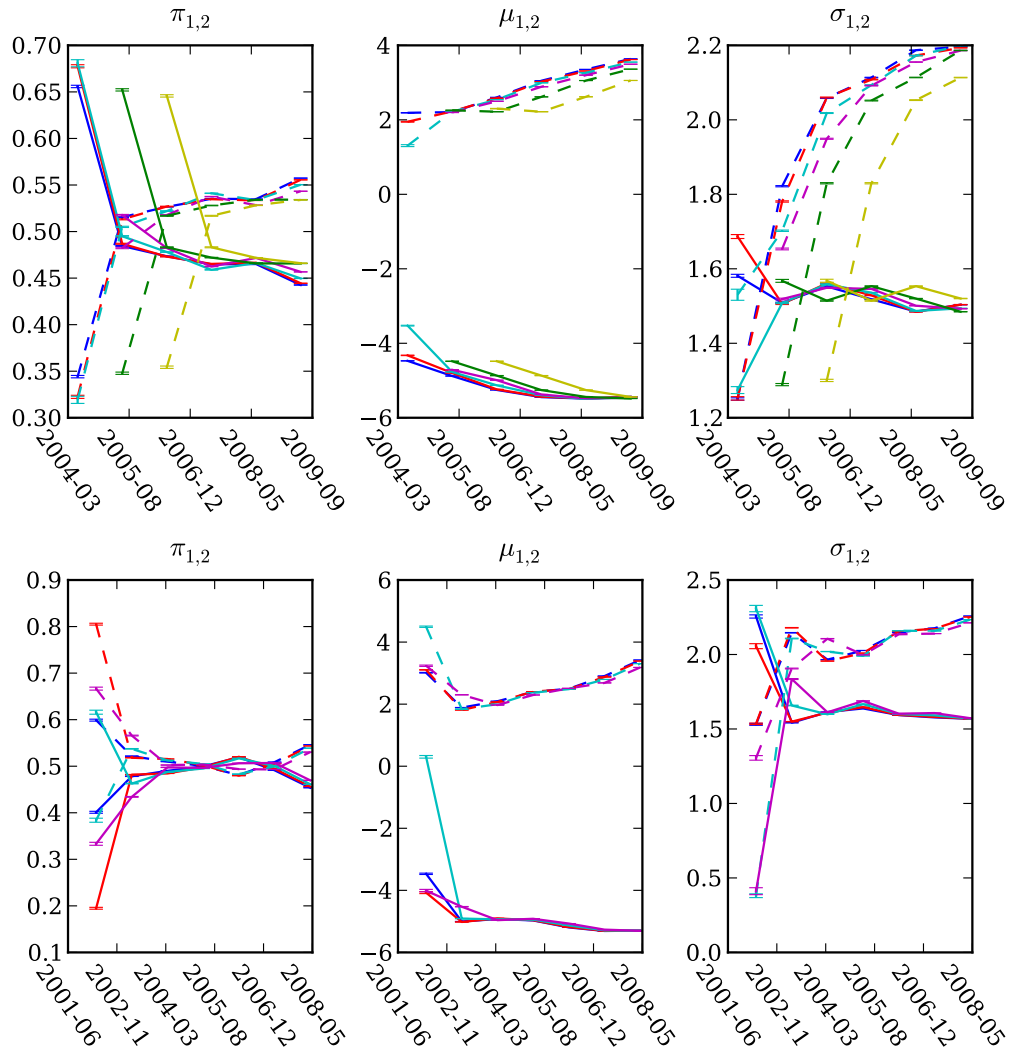


Figure 3.9. Evolution of estimated parameters. Top: Portuguese Wikipedia. Bottom: English Wikipedia. See 3.7 for explanation of symbols.

to play a huge role in shaping participation in a peer production community. As a matter of fact, we see in our analysis that different Wikipedia, thus using the same piece of software, bear different proportions between short-term and long-term users.

This observation strongly motivates the need to take into account the dynamics of social and cultural factors when modeling user participation. The next chapters are devoted to developing a model of peer production that attempts to explain user participation as a process of cultural formation.

Chapter 4

Formation of peer production communities

*Stephane: [Shows 3-D glasses]
You can see real life in 3-D.
Stéphanie: Isn't life already in
3-D?*

“The Science of Sleep” (2006)

4.1 Group formation in online communities

In this chapter we introduce our model of peer production. In the previous chapter we found that an important feature of user activity lifespan in Wikipedia is its multi-modality. A multi-modal distribution suggests, at the very least, the presence of different classes of users. For example, we could imagine that users belong to two different classes, based on their activity lifespan, and ask ourselves what makes a user belong to a particular class. In other words, what accounts for the structure we see within the community of contributors? From a sociological standpoint this amounts asking under what conditions does a core group of long-standing contributors emerge out of a broader population.

As we have already seen, social groups are denoted by their own cultural traits, and online communities are no different in this respect. An example we have already seen from Wikipedia is the adherence to policies such as the Neutral Point of View (NPOV): users of Wikipedia are required to follow this and several other policies, or their contributions will be rejected by the community. There are several other examples of normative behaviors in Wikipedia. Most are

covered by policies, but many are not (cf. Reagle [2007b]), even though they are still documented in the wiki under the form of essays or proposals.¹

Thus, like any real-world social group, a peer production community also exhibits a certain number of norms, established beliefs, shared opinions and, in general, cultural traits, that define it at the social level, and to which any newcomer must comply in order to become a full member of it. In particular, peer production communities have norms related to the production of the digital common around which they are formed. What should be contributed, and how? How to interact with other peers when collaborating? In this chapter we will describe an agent-based model that answers – or at least attempts to answer – these questions.

As we already mentioned in the introduction (see Chapter 1), participation in online communities of peer production is associated with several factors. For example, the already-cited Nov [2007] surveyed Wikipedians asking them what are the most important factors for their participation. The study found that intrinsic motivation, reciprocity, gratification (e.g. fun), a sense of community and extrinsic rewards all correlate positively with an increased level of participation. Correlational findings are difficult to operationalize because they do not necessarily imply causation. Moreover, knowing what factors cause higher levels of participation does not translate easily into knowing what causes users to withdraw from active participation. By means of our model we can test assumptions about user participation and withdrawal, and see if they produce meaningful patterns of user participation.

Motivated by the previous discussion about internal community norms, we can make two assumptions. The first assumption of our model is that a user facing many rejections from his peers will be more likely to withdraw from active participation. The second is that the attitude towards these normative behaviors may change as a consequence of work: by reading the contributions of other members a user might learn the style of writing required by the encyclopedia, for example. Or, by seeing other users resolve disputes by using a civil tone and by assuming other people acted in good faith, he might decide to imitate such behavior.² The way this process of adoption of cultural traits happens is, generally speaking, dictated by the bounded confidence rule we saw in 2.3.3.

¹With this we do not mean to say that formal policies need to be coincident with informal behaviors. In fact, in any organization there are often cases of policy disconnect: what is written is not what is being actually enforced. For an analysis of policy disconnects in Wikipedia, see the so-called “Requests for Adminship” votes analyzed by Burke and Kraut [2008].

²These two latter examples are in fact covered by two official policies. These are respectively: “Civility”, and “Assume Good Faith”.

We thus have two simple mechanisms that account for the formation of a social group. In the rest of the chapter we describe our model more precisely and show its main phenomenological features.

A note on terminology is needed before getting into the description of our peer production model. While we make explicit use of the terminology of wiki platforms (e.g. “users” who “edit pages”) we stress that ours is a general model of group formation in a dynamic bipartite environment, and not merely a description of a wiki platform.

4.2 Model description

Let us consider a dynamic population of users engaged in editing a growing collection of wiki pages. In order to model the evolution of such system, we take into account four types of events. These are:

- A new user joins the community.
- A new page is created by some user.
- A user leaves the project and becomes permanently inactive.
- A user modifies a page.

These four events cover the basic aspects related to peer production that we are interested in.³ As a consequence, our model is composed of several distinct parts, or sub-models, but there is not a 1:1 mapping between model parts and events. We now describe them, and later will discuss how they link to the observed behavior of the model.

4.2.1 Dynamics of cultural traits

We start by describing what happens when a user modifies a page. As we have said, pages are denoted by a certain number of cultural traits, or features, upon which users can find themselves in agreement or not. Features could be opinions,

³In principle we ought to include the deletion of content too. However, most wikis do not allow ordinary users to perform such actions; moreover, those wikis, like Wikipedia, in which page deletions are routinely made, usually delegate the decision to some form of collective deliberation process (Taraborelli and Ciampaglia [2010]). Modeling these discussions is thus beyond the scope of the current work.

norms, beliefs, etc. In this model, we choose to model them as continuous variables. The space in which these cultural features lie is often called the *opinion space* and its definition may reflect different properties of the cultural features and of their dynamics (cf. Lorenz [2007b]). To keep things simple, we consider the elementary unidimensional case, that is, the state of an agent – be it a page or a user – is a scalar number in the interval $[0, 1]$. We denote with $x(t)$ the state of a generic user at time t and with $y(t)$ the state of a generic page.

The bounded confidence rule describes what happens when a user edits a page and, in particular, it captures the dynamics of attitude change in the user. Let us imagine that at time t a user edits a page. If $|x(t) - y(t)| < \varepsilon$ then we update the state of both the page and the user in the following way:

$$x(t) \leftarrow x(t) + \mu(y(t) - x(t)) \quad (4.1)$$

$$y(t) \leftarrow y(t) + \mu(x(t) - y(t)) \quad (4.2)$$

where $\mu \in [0, 1/2]$ is the speed (or uncertainty) parameter and $\varepsilon \in [0, 1]$ is called the confidence parameters (Lorenz [2007b]). In the classic bounded confidence rule, if the condition on the distance between $x(t)$ and $y(t)$ is not met, nothing happens. This is not realistic in our setting, because often users need to deal with a page even though they do not necessarily “agree” with it – for example when fighting vandalism. Most wiki engines – including the Wikipedia software – allow users to undo a contribution with just one click of the mouse; actions of this kind are called rollbacks or reverts. If $|x(t) - y(t)| \geq \varepsilon$, we thus allow the sole (4.2) to take place with probability r . The parameter r represents how likely it is that a user will revert the contribution of someone else, in the event of a disagreement upon their cultural traits.

4.2.2 Editing model

We have described how, when a user interacts with a page, cultural traits change, but not yet how this kind of events happens in the first place.

We should first make some assumptions about how this process happens. The first assumption is that any user chooses to perform his contributions independently of the actions of his peers. At first this might seem a gross simplification, since many times users edit a page because somebody else modified it.⁴ How-

⁴This happens, for example, in the case of editorial conflicts between editors, or when vandalism happens (Kittur, Suh, Pendleton and Chi [2007]; Viegas, Wattenberg and Dave [2004]; Viegas et al. [2007]; Priedhorsky et al. [2007]).

ever, what we are really trying to model here is when users decide to act and start an online session of editing. It is not difficult to imagine that this happens for several factors that are largely independent of what other users did in the past.

Taking this assumption for granted, how do we select users and pages for interaction? Let us consider a user u . In principle, our model should tell us what the probability $P_u(t, p)$ is that, at time t , this user modifies page p . We can always write that:

$$P_u(t, p) = P_u(t)P_u(p | t) \quad (4.3)$$

The first factor on the right hand side of (4.3) gives us the probability that an editor u activates at time t to perform an edit. The second factor specifies, given the activation at time t , what page p will be selected – basically what page the user will choose to edit. Regarding the probability $P_u(t)$, we specify two different editing models, which we describe now.

Homogeneous edit process

The simplest model for $P_u(t)$ is to assume that any time is equally likely for an individual to perform an edit. This means that the editing behavior of a user is modeled as a Poisson process with homogeneous rate λ_e .

This rate should in principle depend on the user, since different people have different activity levels, but since we are interested in modeling the overall span of the activity of users, and not their edits count, we can accept a further simplification, and thus assume that all users perform edits at the same rate λ_e . We used this simpler model of editing in chapter 5, where we want to understand how important the parameters related to the dynamics of opinions are – the bounded confidence parameters.

Edit process with Poissonian cascades

A homogeneous process is not capable of capturing one essential aspect of the real editing activity of users of an online community – *burstiness*. In fact, a wide range of user activity logs (e.g. email sending, trading, phone calls, SMSs, etc.) show that events tend to happen in rapid sequence, and that clusters of events are followed by long periods of inactivity (cf. Barabási [2005]).

To model this, we consider a model of edit cascades. As before, we assume that a user activates to perform an edit with a constant rate λ_a . Once he is active, he performs, on average, N_a additional edits with rate λ_e . The burstiness

effect is thus obtained by assuming that the two rates differs significantly, i.e., $\lambda_e \gg \lambda_a$, by at least an order of magnitude. This model is more realistic because it better captures the lifespan of short-term users. We used it in chapter 6 when comparing model output with empirical data.

Page selection

Having seen how we compute $P_u(t)$, we now need to specify how we model $P_u(p|t)$, the probability that a user selects a given page for editing. The first assumption we make is that the probability of selecting a page does not depend on who is selecting it, that is, $P_u(p|t) = P(p|t)$.

The second assumption is that different pages can reflect different topics and hence, based on their popularity, receive different levels of attention from users. We employ a simple reinforcement mechanism called preferential attachment (or PA) to model this (cf. Barabási and Albert [1999]). Let $c_p \geq 0$ be a constant. If m_t is the number of edits a page has received up to time t , then the probability of it being selected at that time will be proportional to $m_t + c_p$:

$$P(p|t) \propto m_t + c_p \quad (4.4)$$

When $c_p \rightarrow \infty$, pages will be chosen for editing with uniform distribution, regardless of the number of edits they have received. Hence, we can study the impact of content popularity in user participation by setting c_p to a small or large value.

4.2.3 Population dynamics

We now turn to describe the three kinds of events that are responsible for the dynamics of the whole population of users. These are: the arrival of a new user; the creation of a new page; the departure of a user.

Let us start by considering users. For new arrivals, we simply consider a homogeneous input rate of new users ρ_u .⁵ The state of each new user is chosen at random within the interval $[0, 1]$.

⁵Of course the probability of joining a community such as Wikipedia is highly dependent on exogenous factors such as the popularity of the project and on societal, organizational, and cultural trends. Similarly, for the process of page creation, we might imagine that the probability that a user creates a page will depend on several external factors that are hard to quantify, for example how many topics are still waiting to be covered at a given stage of the development of the encyclopedia. We study the inhomogeneity of the edit activity rate in chapter 7.

In the case of departures, we consider instead an inhomogeneous departure rate. Let us consider a generic user at time t and let us denote with n_t the number of edits he (or she) did up to t , and with s_t the number of these edits that resulted in the application of (4.1). Let $c_s \geq 0$ be a constant and $r(t)$ be the ratio:

$$r(t) = \frac{s_t + c_s}{n_t + c_s} \quad (4.5)$$

The rate of departure $\lambda_d(t)$ is then defined as:

$$\lambda_d(t) = \frac{r(t)}{\tau_0} + \frac{1 - r(t)}{\tau_1} \quad (4.6)$$

with $\tau_0 \gg \tau_1$ time scale parameters. Depending on the value of $r(t)$, the expected activity lifespan $\langle \tau \rangle$ will interpolate between two values: $\langle \tau \rangle = \tau_0$ (long lifespan) for $r(t) = 1$, $\langle \tau \rangle = \tau_1$ if $r(t) = 0$ (short lifespan). If $c_s \rightarrow \infty$, we recover a homogeneous process with rate τ_0^{-1} , so we can set c_s to control how sensitive the departure rate is to unsuccessful interactions.

Let us now consider pages, and in particular the creation of new pages. We model this event by considering a constant rate ρ_p at which new pages are created. The creator of a new page is chosen at random among the existing users. Whenever a new page is created, its cultural state y is equal to that of the creator x .

4.3 Model implementation

We can simulate our model using an exact simulation algorithm by Gillespie [1977]. For each existing user there are two Poisson processes at work: one for editing, and one for departure. In principle the departure process has an inhomogeneous activity rate, which means that we need to update the state of the system (in particular the number of edits n_t and the number of successes s_t) every time a new edit is performed. This is true also for the processes of user arrival and page creation.

Thus, at any time t , we have a compound process, whose global activity rate can be computed from the activity rates of each user (edit and departure), plus the global rates of user arrival and page creation. It is worth noting that a similar approach is generic for any kind of agent-based model, not just for our peer production model (for example cf. Vancheri et al. [2008]).

4.4 Simulation of cultural dynamics

Even though we devote the next two chapters to the analysis of the model, we can spend some time now describing the behavior of the model. In particular, it would be interesting to understand if the dynamics of the model by Deffuant et al. features any qualitative difference when applied to a population whose individuals are subjected to (4.6) – as it is in our case. We thus performed some exploratory simulation to better understand this point.

Figure 4.1 shows five different realizations of the evolution of the distribution of the cultural state, or opinion, of a population of users. Each red line represents the path of the state of an agent. The “jumps” we can see in the plots correspond to editing events. The simulations were obtained by progressively raising the value of the confidence parameter ε , while all other parameters were kept constant. In particular, ε takes five evenly spaced values in the interval $[0, 0.2]$.

The plots show that for low values of ε we have a fragmented state, in accordance with the traditional Deffuant model (fig. 4.1a). At this stage the average activity lifespan is rather short, hence the turnover of users prevents the emergence of stable groups, or clusters. As ε grows, the average lifespan of users grows as well (fig. 4.1b).

However, if we raise ε a bit more, we see the occurrence of a new phenomenon. In fig. 4.1c, after an initial fragmented phase, approximately after $t = 5,000$ iterations⁶ a central cluster of users emerges. These users all share the same cultural trait within distance ε . The cluster forms because it is able to self-sustain, namely it reaches a critical mass of users who collectively control the majority of pages. Looking at fig. 4.1c, we can note that this phenomenon is accompanied by a sudden decrease of fluctuations and disappearance of minority fragments in the rest of the opinion space. The average lifetime in the central cluster becomes much longer than outside. A core group of users, or majority, has thus emerged.

However, at this stage the core group is not large enough to be able to sustain itself indefinitely and we can see, approximately at time $t = 15,000$, that fluctuations resume and minority agents reappear at the periphery of the opinion space.

Raising ε even further we see that the emergence of the cluster happens earlier (4.1d), and that the central cluster lasts for a longer period. For $\varepsilon = 0.2$, no initial fragmented phase can be discerned at all (4.1e). We can also see that

⁶In these simulations we use discrete time steps, instead of the point process explained before.

the central cluster is able to attract users whose initial opinion is close to it. Users with more extreme opinions fail to be incorporated in this group.

Interestingly, users with “extreme” cultural traits never manage to survive long enough to form a lasting minority group. A possible explanation is that the page selection mechanism prevents this from happening. In fact, the preferential attachment rule (4.4) applies regardless of the value x of the cultural variable. The influx of new pages is not a remedy against this phenomenon, and we can illustrate this point with a *reductio ad absurdum*: let us imagine that a group of users with extreme opinions (in the sense of being farther than ε from the majority) has emerged thanks to, for example, a random sequence of user arrivals and page creations. Users in this group score enough successes and thus their expected lifetime is high. Of course if they edit a page outside of their pool, they will expose themselves to the majority view, experience “failure”, and thus shorten their lifespan. So let us imagine that, again by a random fluctuation, these users keep on editing their own pool of pages. But eventually these pages will accrue enough edits and become popular also for the majority group, which will start selecting them, thus overtaking the minority.

4.5 Discussion

As evidenced throughout the chapter, the present model makes several assumptions. It is beneficial to review them all here in order to highlight limitations that could be addressed in future works.

- **Deletion of content.** We assume that content deletion is not an important interaction for the determination of the activity patterns. In peer-production communities, users are generally neither required nor allowed to delete contents forming the digital common, and thus this assumption is perfectly fine for the vast majority of peer production communities. In some cases, decision about content inclusion are either left to a small support team, namely for copyright infringements and other cases of unlawful content. Sometimes inclusion decision are crowdsourced themselves; in the case of Wikipedia there is an articulated deletion process. In these cases it seems plausible to consider whether divergences of views and other forms of conflict might cause users to withdraw from participation.
- **Uncorrelation activity.** We assume that edits occur independently of the activity of other editors and other stimuli. While this is clearly a simplification, we should note that several examples of Internet activity are not due

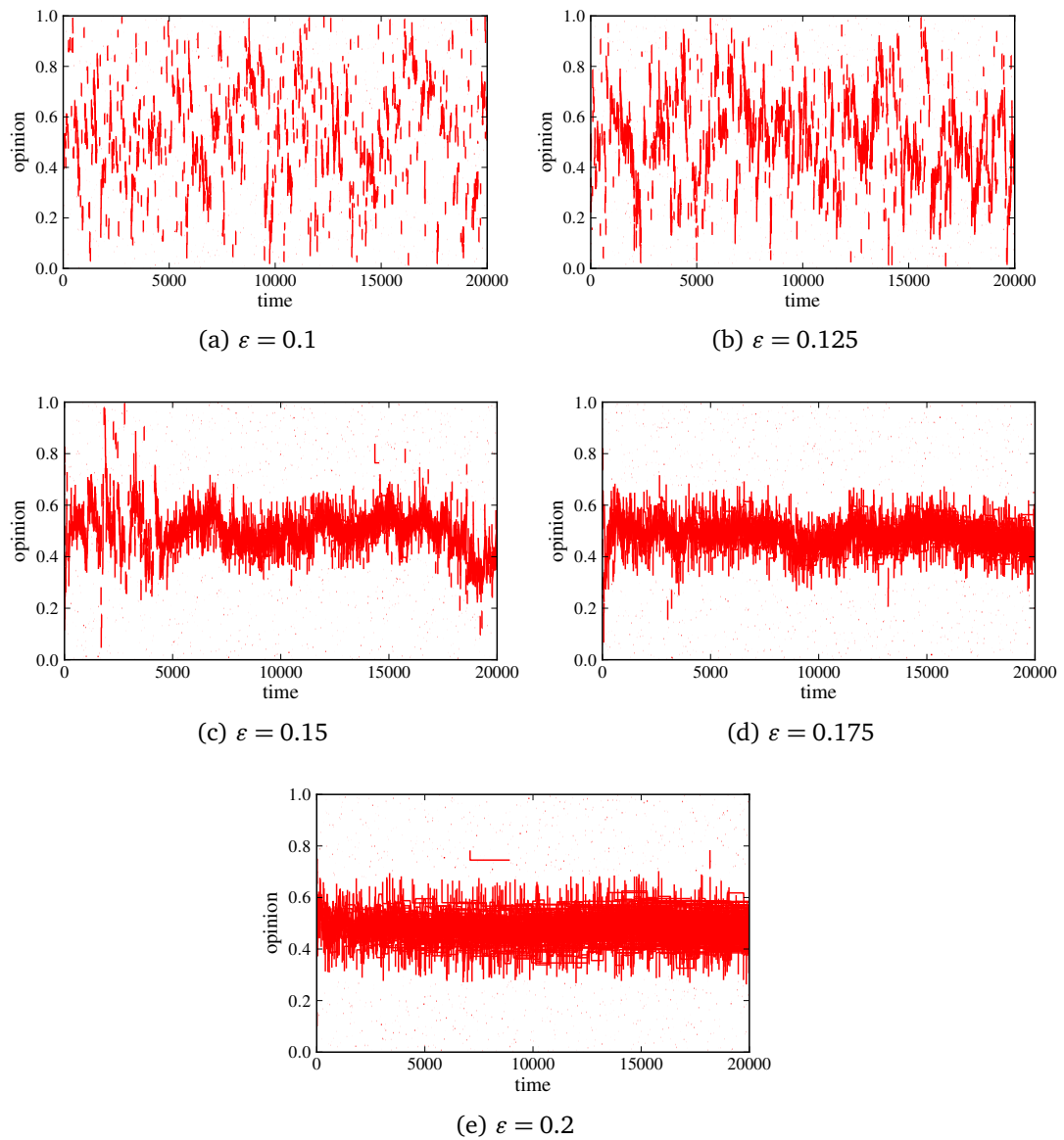


Figure 4.1. Dynamics of group formation. Each red line represents the opinion of a single user over time.

to communication or other forms of dependence between individuals. For example, Rafaeli and Ariel [2008] point out that Wikipedia participation may also have a routine component. On the other hand, e-mails, IM notifications, and similar technical features (for example Wikipedia watchlists) might introduce a degree of inter-personal correlation that thus need to be taken into account.

- **Homogeneous editing rates.** Editing activity was assumed to be homogeneous over time and across different users. As we have already noted, for the scope of measuring the long-term activity lifespan of users, the degree to which our model reproduces the day-to-day activity patterns of users is not important. Measurements at shorter time scales are instead sensitive to the fine-grained activity patterns, and that is why, for the sake of calibration, we introduce a model with Poissonian cascades. Two other aspects could be improved further. In fact, we did not account for heterogeneity of activity rates neither at the population level, nor during the lifespan of individuals. Of course, before any modeling takes place, a sensible approach would be first to understand the patterns at an empirical level. Chapter 7 is devoted to this task. activity at an empirical
- **User inactivity.** In our model, users become permanently inactive. In reality, people could take a long break and resume activity later. As with the assumption of homogeneous editing rates above, this limitation of the model can be addressed only after looking at the empirical patterns of user activity.
- **User editing interests.** In the page selection model we assumed that $P_u(p|t) = P(p|t)$, that is, all users choose which page to edit according to the same model. In reality, people have different tastes and interests and thus it makes sense to ask how to model these user profiles. For example introducing memory effects in a similar way to Crandall et al. [2008], who used a complex network urn model that would take into account the effect of social influence.
- **Page popularity.** Pages keep on accruing popularity depending on the amount of attention they received in the past, which we model by considering the number of past edits. In reality, collective attention is known to fade with a characteristic time scale (Wu and Huberman [2007]), and also not all topics have the same intrinsic popularity. Thus individual pages could be assigned an initial fitness value c_p , and fitness could be made to

decay over time in order to reflect the empirical observations about collective attention.

- **Constant rate of new users and pages.** For the sake of simplicity we assume that new users arrive at a constant rate. Model predictions are likely to improve, if a realistic influx rate $\rho_u(t)$ would be employed. A similar argument holds for the rate of page creations ρ_p .
- **Minorities.** As we saw, the model features two distinct regimes, a fragmented state and a full consensus (i.e. monocluster) state. Minorities play an important role in intergroup dynamics, and indeed the original Defuant model features group polarization for certain values of the parameter ε . As we explained before, this limitation is due to the page selection model and thus could be easily addressed later. In fact, at this stage of research we are interested in looking at the emergence of a core group and thus the current model is perfectly amenable for investigation against empirical data.

Chapter 5

Factor screening and sensitivity analysis

— *Sensitivity analysis for
modellers?*
— *Would you go to an
orthopaedist who didn't use
X-rays?*

Fürbinger [1996]

5.1 Introduction

Computational and mathematical models usually have a number of parameters in their specification. Parameters, which can be scalar, vector-valued, or functional (like a time series), are meant to affect in some way the output of their model. In this sense all models, even those that are not analytic – like agent-based simulation models – can be thought of as being functions of their parameters.

Of course this mapping can be either deterministic or stochastic; besides this distinction, in both cases one might be interested in quantifying how much of the “variability” of the response of the model can be apportioned to each of its parameters. Sensitivity analysis (SA) is a set of statistical techniques commonly used to answer this question.

What is sensitivity analysis useful for? One application is for factor screening: we want to know which parameters, or factors, are the most important in

accounting for the variability of response of the model (Saltelli et al. [2004]). If the model has many parameters, this information can be useful for guiding, at a later stage, the process of model calibration or even for performing a more informed data collection.

But this is not all. If we move in the context of computational (agent-based) models, we see that one has usually to make several assumptions when specifying the microscopic behavior of agents. In fact, it is legitimate to ask how much the overall collective behavior of agents is robust to changes in the set of rules or assumptions. With n rules, each with k possible alternatives, this amounts to exploring n^k alternative models, which becomes unfeasible quickly. In this situation, Ellner and Guckenheimer [2006] propose to use sensitivity analysis as an instrument for escaping this problem, which is commonly known as the “curse of dimensionality”.

In this chapter we detail some methods for performing SA and report the results of the study of factor screening for our proposed model of peer production.

5.2 Methods

5.2.1 Local sensitivity analysis

Roughly speaking, the concept of sensitivity of a function has to do with how much the function changes when one of its inputs changes by a given amount. Let us start by considering a deterministic model with d parameters, which we organize into a vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$. The model produces an output, for simplicity a scalar y . We can thus think of this model as a function that maps values of the parameters to values of the output, i.e. $y = f(\boldsymbol{\theta})$, so a measure of the importance of θ_i can be the increase (or decrease) in the output y for an infinitesimal increase of θ_i , i.e. the partial derivative in θ_i :

$$s_i(\boldsymbol{\theta}) = \frac{\partial f}{\partial \theta_i} \quad (5.1)$$

If we define a sensitivity index in this way, we run into a major problem: each θ_i has a different scale and unit of measure; this makes s_i not comparable with another s_j . We can overcome this issue if, instead of a plain derivative, we consider the ratio between the fractional changes of the two variables y and θ_i , that is:

$$\hat{s}_i(\boldsymbol{\theta}) = \frac{\partial f}{\partial \theta_i} \cdot \frac{\theta_i}{f(\boldsymbol{\theta})} \quad (5.2)$$

This quantity, now dimensionless and normalized, can be used to rank all parameters. In fact, Ellner and Guckenheimer [2006] point out that, for a matrix model of population dynamics, (5.2) is just its *elasticity*. The term “elasticity” is more common than “sensitivity” among economists and in the literature on dynamical systems.

It is important to point out, however, that the sensitivity/elasticity in (5.2) is only a *local* quantity. If we want to get a global picture of the importance of its parameters, we have to evaluate our model at several locations. On the other hand, if we want to analyze our model for a given set of parameters, we can implement equation (5.2) with Newton’s finite difference formula at the expense of $2d$ additional evaluations of y , for any given θ .

Another shortcoming of a sensitivity analysis based on equation (5.2) is that index s_i does not account for the so-called *interaction* effects between parameters. Imagine that y is rather insensitive to (fractional) changes of θ_i but that changing θ_i suddenly makes y more sensitive to another parameter θ_j . We would like to account for this indirect effect of θ_i on y when ranking the relative importance of all parameters. But equation (5.2) does not contain any term related to other θ s, so it cannot reflect this.

We now move to the subject of *global* sensitivity analysis and see how the methods proposed in that context address the issues we have just seen.

5.2.2 Three approaches to global sensitivity analysis

In the previous section we introduced some basic concepts of sensitivity analysis using a deterministic function $y = f(\theta)$. Before describing global sensitivity analysis, we have to move from this deterministic setting to a more realistic one.

Models that seek to capture aspects of reality take into account various forms of uncertainty and noise in the processes they represent. This changes the mathematical treatment considerably: instead of a deterministic response variable y , one has to consider a random variable Y as output of the model; instead of evaluating $f(\theta)$, one has to consider its expected value $E(Y | \theta)$. In practice we record several observations of the model response and use them to estimate the model response for a given θ .

Randomness does not affect only the response of the model, though. In many cases we specify a model without knowing the “true” – or best – values of its parameters θ . Sometimes we are lucky and have access to direct field measurements of them. At some other times we are less so and have just point estimates or credible regions (in a Bayesian context) coming from other modeling studies. At the bottom of such a ranking by desirability we just have a range that we

think might be more or less plausible. In some cases this is perfectly legitimate – if the parameter has the meaning of a probability, we know by definition that it lies in $[0, 1]$ – in some other cases not so much.

In all these three scenarios there is a source of uncertainty that we cannot rule out and that derives from our lack of knowledge of the true value of the parameters of the processes we are modeling. However, uncertainty is not just a matter of incomplete knowledge. In fact, even if we were to know the “true” value of the parameters, we would still not be able to rule out randomness completely.

Think for example of some engineering process that we want to model. Engineers usually distinguish between *environmental factors*, which are subject to noise and cannot be controlled, and *process factors*, which are instead those under the engineer’s control. Santner et al. [2003] give the example of the amounts of the various ingredients in the recipe of a cake (process factors) versus the temperature of the kitchen’s oven (environmental factor) we use to bake the cake with. It is clear from this distinction that even in the absence of any source of noise due to the environment, we would still not be able to rule out the uncertainty in how we set the process factors.

So we are left with no other option than to consider the input factors θ to be random variables as well. In particular, if they are independent random variable, we can write the joint probability distribution $P(\theta)$ from the marginals $P_i(\theta_i)$, for $i = 1 \dots d$. This is usually a fairly reasonable assumption – unless there is some reason to suspect the opposite. As we said before, we can take these marginals to reflect any prior knowledge we have on the distribution of parameters. In the following sections we will use the uniform distribution, but in principle anything could do, for example the Gaussian distribution.

Thus, from now on, we will look at a stochastic response model. Having introduced the general setting for global sensitivity analysis, we now see three methods that are commonly used in the literature to perform it.

Partial correlations

Correlation is perhaps the simplest way to look at the effect of a parameter on the response of a stochastic model, but it also has some limitations, namely that it can capture only linear effects and that it may give misleading results when applied to non-monotonic models. The Pearson correlation coefficient between the i -th parameter and output y does account for *linear* effects and is defined as:

$$\rho_i = \text{corr}(\theta_i, Y) = \frac{E[(\theta_i - E[\theta_i])(Y - E[Y])]}{\sigma_{\theta_i} \sigma_Y} \quad (5.3)$$

where σ_X is the standard deviation of random variable X . However, the correlation coefficient has the problem that it does not control for the effect of other parameters. Instead, one uses the *partial* correlation coefficient of θ_i and Y given $\boldsymbol{\theta}_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_d)$, i.e. $\rho_{i \cdot -i} = \text{corr}(\theta_i, Y | \boldsymbol{\theta}_{-i})$ (note the dot notation).

The partial correlation coefficient is easily computable since it is the correlation coefficient of the residuals r_Y and r_{θ_i} , obtained respectively by regressing Y and θ_i on the set of conditioning parameters $\boldsymbol{\theta}_{-i}$.

Standardized regression coefficients

Another option for quantifying linear effects is to standardize variables Y and $\boldsymbol{\theta}$

$$\tilde{Y} = \frac{Y - \bar{Y}}{\sigma_Y}, \quad \tilde{\theta}_i = \frac{\theta_i - \bar{\theta}_i}{\sigma_{\theta_i}} \quad \forall i = 1, \dots, d, \quad (5.4)$$

and to fit a linear regression model

$$\tilde{Y} = \beta_0 + \beta_1 \tilde{\theta}_1 + \dots + \beta_d \tilde{\theta}_d \quad (5.5)$$

so that the regression coefficients are standardized too. Saltelli et al. [2004] points out that, in a linear model of the form $Y = \langle \mathbf{C}, \boldsymbol{\theta} \rangle$, with independent parameters and having Gaussian distributions, then s_i^σ , the standardized version of the sensitivity index s_i of (5.1),

$$s_i^\sigma = \frac{\sigma_{\theta_i}}{\sigma_Y} \frac{\partial Y}{\partial \theta_i} = \frac{\sigma_{\theta_i}}{\sigma_Y} C_i \quad (5.6)$$

can be directly computed by means of the standardized regression coefficient β_i . Moreover, since in this specific case (but not in general) the squared output variance is the sum of the squared input variances,

$$\sigma_Y = \sqrt{C_1 \sigma_{\theta_1}^2 + \dots + C_d \sigma_{\theta_d}^2} \quad (5.7)$$

then the regression coefficients give the fraction of output variance each parameter accounts for.

Unfortunately, (5.7) holds only in the (limited) example of a linear model

with independent Gaussian inputs. For any other model more complicated than this, we lose this nice interpretation.

In summary, the caveat for both standardized regression coefficients and partial correlation coefficients is that they characterize only linear and monotonic effects of the parameters on the response. What is nice about these methods is that we can *always* compute them, but we have to keep in mind that if our model is non-linear, then they might provide a distorted picture of the importance of the parameters, even when the coefficient of determination R^2 is high.

The third and last method we will see is based on the decomposition of the output variance, and is not affected by these problems.

Variance decomposition

This method was proposed by Sobol' [2001] and is based on the analysis of variance (ANOVA). The idea is to decompose the variance of the output in several components that are attributable to independent factors, in our case the parameters of the model.

Let us assume that the space of parameters is $[0, 1]^d$. Sobol' proposes to write the output Y as:

$$Y(\theta_1, \dots, \theta_d) = Y_0 + \sum_{i=1}^d Y_i(\theta_i) + \sum_{1 \leq i < j \leq d} Y_{i,j}(\theta_i, \theta_j) + Y_{1,2,\dots,d}(\theta_1, \theta_2, \dots, \theta_d) \quad (5.8)$$

and shows that this decomposition is unique under two assumptions. The first is that, $\forall m = 1, \dots, d$ and for each $\{i_1, \dots, i_m\} \subseteq \{1, \dots, d\}$:

$$\int_0^1 Y_{i_1, \dots, i_m}(\theta_{i_1}, \dots, \theta_{i_m}) d\theta_{i_k} = 0, \quad \text{for } k = 1, \dots, m. \quad (5.9)$$

The second is that all summands are orthogonal, that is, for any two subsets of indices $A \neq B$:

$$\int_{[0,1]^d} Y_A \cdot Y_B d\theta_1 \cdots d\theta_d = 0 \quad (5.10)$$

In the above, $Y_0 = E[Y]$ is the expected value of Y , the term $Y_i(\theta_i)$ is called the *main* effect of parameter θ_i , and the term $Y_{i,j}(\theta_i, \theta_j)$ is the *interaction* effect between the i -th and j -th parameters ($i \neq j$). In general, if $A = \{i_1, \dots, i_m\}$ is

a subset of m indices, then $Y_A(\theta_{i_1}, \dots, \theta_{i_m})$ is the interaction effect of order m , or m -way interaction effect, of A . Each summand is computable from suitable integrals. For example the main effect of Y_i is:

$$Y_i(\theta_i) = \int_0^1 \dots \int_0^1 Y(\theta_1, \dots, \theta_d) d\boldsymbol{\theta}_{\neg i} - Y_0 \quad (5.11)$$

where with $\boldsymbol{\theta}_{\neg i}$ when mean the reduced parameter vector obtained by considering all parameters except θ_i . Similar formulas can be obtained for higher order effects. Let us now consider the variances of the summands in (5.8): $\sigma_i^2, \sigma_{i,j}^2, \dots$, etc.; we can decompose σ^2 , the total variance of Y , as:

$$\sigma^2 = \sum_{i=1}^d \sigma_i^2 + \sum_{1 \leq i < j \leq d} \sigma_{i,j}^2 + \dots + \sigma_{1,2,\dots,d}^2 \quad (5.12)$$

The sensitivity indices proposed by Sobol' [2001] are thus obtained by standardizing all summands of (5.12), obtaining:

$$1 = \sum_{i=1}^d M_i + \sum_{1 \leq i < j \leq d} C_{i,j} + \dots + C_{1,2,\dots,d} \quad (5.13)$$

M_i is the *main sensitivity index* of parameter θ_i , $C_{i,j}$ is the *two-way interaction index* between θ_i and θ_j , etc. Two quantities are of interest for assessing the importance of a parameter: the already cited main sensitivity index M_i ; and the *total interaction index* T_i , which is defined as the sum of all terms that involve parameter θ_i :

$$T_i = \sum_{j \neq i} C_{i,j} + \sum_{\substack{1 \leq j < k \leq d \\ j, k \neq i}} C_{i,j,k} + \dots + C_{1,2,\dots,d} \quad (5.14)$$

How to compute M_i and T_i ? Jansen et al. [1994] proposes a method based on resampling. Let us imagine splitting the vector of parameters in θ_i and $\boldsymbol{\theta}_{\neg i}$ and to look at the difference $Y(\theta_i, \boldsymbol{\theta}_{\neg i}) - Y(\theta_i, \boldsymbol{\theta}'_{\neg i})$, where $\boldsymbol{\theta}_{\neg i}$ and $\boldsymbol{\theta}'_{\neg i}$ denote two independent random draws with joint distribution on all parameters except θ_i (remember that the factors are independent). If we apply the decomposition in (5.8), then we see that all terms except the i -th main effect Y_i are duplicated and cancel out. By (5.11) its variance is then:

$$E \left[Y(\theta_i, \boldsymbol{\theta}_{\neg i}) - Y(\theta_i, \boldsymbol{\theta}'_{\neg i}) \right]^2 = 2(\sigma^2 - \sigma_i^2) \quad (5.15)$$

and if we normalize we finally obtain:

$$M_i = 1 - \frac{1}{2\sigma^2} E \left[Y(\theta_i, \boldsymbol{\theta}_{-i}) - Y(\theta_i, \boldsymbol{\theta}'_{-i}) \right]^2 \quad (5.16)$$

A similar argument, this time considering the difference $Y(\theta_i, \boldsymbol{\theta}_{-i}) - Y(\theta'_i, \boldsymbol{\theta}_{-i})$, leads to the following formula for the total interaction index T_i :

$$T_i = \frac{1}{2\sigma^2} E \left[Y(\theta_i, \boldsymbol{\theta}_{-i}) - Y(\theta'_i, \boldsymbol{\theta}_{-i}) \right]^2 \quad (5.17)$$

In the following section we will see a sampling method that let us evaluate (5.16) and (5.17) in an efficient way.

5.2.3 Experimental design

The methods we have seen so far require computing Monte Carlo estimates of Y over the full space of parameters. This can be done either through direct simulation or by resorting to *surrogate* models that approximate the response of our original model. In both cases, one question still has to be answered: where shall we evaluate our model to produce estimates of Y ? We know from standard Monte Carlo that the straightforward way is of course to sample from the marginal distributions P_i a series of points $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(n)}$ and obtain observations Y_1, \dots, Y_n . Usually these observations are themselves averages of multiple model realizations of Y for a given $\boldsymbol{\theta}$. We will call these averaged observations *model runs*.

Approximating an integral via Monte Carlo simulation with uniform sampling gives us an error estimate that goes to zero like $(\sqrt{n})^{-1}$, as $n \rightarrow \infty$. Are there more efficient ways to direct the exploration of the parameter space, also from the computational point of view?

Experimental designs respond to such a question. One of the first alternatives to uniform sampling was the one of Sobol' [2001] who proposed using *quasi-random* sequences. This approach gives error bounds of the order of n^{-1} . In the next section we will see a method based on Latin Squares that has similar error bounds and is very simple to implement.

Latin Hypercube Sampling

Uniform sampling is not the only way to generate inputs for our simulations. McKay [1992] proposes the Latin Hypercube Sampling (LHS), which is defined so that each subinterval of a subdivision in quantiles of the support of a parameter appears exactly once. By “appears” we mean that a representative point is

selected from each interval, for example the midpoint.

Let us see a simplified example. Consider the case in which all marginals P_i are uniform, so that the quantiles are evenly spaced. For any j , we say that the interval (a_j, b_j) is the support of marginal P_j . We subdivide these intervals in n sub-intervals, of equal probability, and consider their midpoints:

$$x_i^{(j)} = \frac{(b_j - a_j)(2i + 1)}{2n} + a_j, \quad \text{for } i = 0, \dots, n - 1. \quad (5.18)$$

To construct a Latin Hypercube sample we build a sample matrix:

$$\Theta = (\theta_1, \dots, \theta_n)$$

but instead of taking $(\Theta)_{ij} = \theta_{ij} = x_i^{(j)}$, we randomly shuffle the coordinates along each dimension. The j -th column of the matrix is thus obtained by taking a permutation $(k_{1,j}, \dots, k_{n,j})$ of the indices $(1, \dots, n)$, so that:

$$\theta_{ij} = x_{k_{i,j}}^{(j)} \quad (5.19)$$

In practice the set of indices $k_{i,j}$ is itself a Latin Square over the alphabet $\{1, \dots, n\}$. A Latin Hypercube design has the property that any projection of its points in $d - 1$ dimensions is still uniformly distributed over the resulting subspace.

How to choose a “good” Latin Hypercube sample? Of course any collection of d permutations of $(1, \dots, n)$ will generate a legal design, but many of them will not spread well across the space of parameters (e.g. take the case of no shuffling at all). A good *space-filling* design should instead minimize the error of the sensitivity indices estimates, which is a highly desirable property.

This problem is solved by taking a *maximin* design. A maximin Latin Hypercube sample (LHS) is a sample that maximizes the minimum distance over its points, that is:

$$\max_{\Theta} \min_{i < i'} \|\theta_i - \theta_{i'}\| \quad (5.20)$$

Winding Stairs Sampling

Equations (5.16) and (5.17) can be implemented straightforwardly by fixing part of the vector of parameters and then running the simulator (or its surrogate). The original method, proposed by Sobol', uses two sample matrices, the second of which is called the *resampling* matrix. Jansen et al. [1994] proposes instead

a simpler method to compute the main and total interaction indices, which is called Winding Stairs (ws) sampling.

The idea of a Winding Stairs sample is to generate input samples by updating one parameter at a time, in a cyclical fashion. Observations of Y are then arranged in a special matrix with r rows (the size of the sample) that simplifies the computation of M_i and T_i .

Let us consider $d = 3$ and $r = 4$. The following matrix forms a Winding Stairs sample:

$$W = \begin{pmatrix} Y(\theta_{11}, \theta_{21}, \theta_{31}) & Y(\theta_{11}, \theta_{22}, \theta_{31}) & Y(\theta_{11}, \theta_{22}, \theta_{32}) \\ Y(\theta_{12}, \theta_{22}, \theta_{32}) & Y(\theta_{12}, \theta_{23}, \theta_{32}) & Y(\theta_{12}, \theta_{23}, \theta_{33}) \\ Y(\theta_{13}, \theta_{24}, \theta_{33}) & Y(\theta_{13}, \theta_{24}, \theta_{34}) & Y(\theta_{13}, \theta_{24}, \theta_{35}) \\ Y(\theta_{14}, \theta_{25}, \theta_{34}) & Y(\theta_{14}, \theta_{25}, \theta_{35}) & Y(\theta_{14}, \theta_{25}, \theta_{36}) \end{pmatrix} \quad (5.21)$$

here θ_{ij} is the j -th sample of the i -th variable. Updates along a row proceed from left to right, moving to the next row each time we reach its end. The first element has parameters $(\theta_{11}, \theta_{21}, \theta_{31})$. On moving to the next element, the first update involves the second parameter, i.e. θ_{21} is updated with θ_{22} . Then the third, etc. At the end, the first parameter is updated upon moving to the next row, and the cycle starts again. In this way a column forms an independent sample of observations, but the same does not hold across different columns.

According to equation (5.16), the main index M_i is estimated as the total variance σ^2 minus half the average squared difference between elements that are separated by $d - 1$ updates. As we said, moving $d - 1$ steps on the right in the design causes us eventually to jump to the next row. Hence for the first parameter differences are taken between the first and the last column. For the second parameter differences are taken between the second and the first column, with elements of the first column shifted by one row, and so on for all other parameters.

Estimation of equation (5.17) proceeds in a similar fashion: the i -th total interaction index T_i is half the average squared difference between pairs of elements of the matrix for which only one input value (the i -th) has changed. These elements are adjacent, i.e. one step on the right and optionally moving to the next row. The first parameter changes its value for the first time at the end of the first row, so the differences are computed between elements of the last and the first column, with elements from the first column shifted down by one row. For all other parameters, instead, the first change of value happens within the first row, so in this case differences are taken over two adjacent columns and

rows are aligned.

5.3 Results

In this section we report the results obtained using the three methods described above. Our goal is to perform a *factor screening* of the most important parameters of our model of peer production – those that account for most of the output variance of the model. This information will be useful in the subsequent calibration phase.

During that phase, in fact, we will try to match the output of our model with field data taken from real online communities, and it would be good to know which parameters can be fixed, for example to some value taken from the literature, and which need instead a careful calibration. Reducing the number of parameters will reduce the computational burden of the model fitting technique, and this is something we would welcome very eagerly.

The output quantity we use in the sensitivity analysis is the average activity lifespan τ . A run of our model will produce a distribution of lifespans for all agents in the simulation, so we will take $\langle \tau \rangle$, the average τ , as a scalar measure of the longevity of users within the peer production community.

We will now see in detail the simulation scenario employed throughout the analysis.

5.3.1 Simulation scenario

Table 5.1 lists all parameters of the model. Two quantities are fixed for the purposes of this analysis: simulation time, and transient time. Two other parameters, the initial number of users N_u and pages N_p are dynamical variables and are determined by means of an initial transient. All remaining parameters are allowed to vary in their own interval. Thus we have an input space with 10 dimensions.

The two activity lifespans τ_0 and τ_1 were chosen to range in non-overlapping intervals corresponding to different time scales, consistent with empirical observations of user participation from Wikipedia (see Ciampaglia and Vancheri [2010]). Simulation time T was chosen consequently; transient time T_0 was set to twice the value of T (see section 5.3.1).

For physical quantities like daily rate of edits (λ_e), new user arrivals (ρ_u), and new page creations (ρ_p) intervals have been chosen looking at plausible values from the public statistics on the Wikipedia project. These statistics are freely

Parameter	Symbol	Values	Unit	Distribution
Const. popularity	c_p	(0, 100)		uniform
Const. successes	c_s	(0, 100)		uniform
Confidence	ε	(0, $1/2$)		uniform
Daily edit rate	λ_e	(1, 20)	day	uniform
Daily rate of pages	ρ_p	(1, 20)	$1/\text{day}$	uniform
Daily rate of users	ρ_u	(1, 20)	$1/\text{day}$	uniform
Initial no. of pages	N_p			see §5.3.1
Initial no. of users	N_u			see §5.3.1
Long activity lifespan	τ_0	(10, 100)	day	uniform
Rollback probability	r	(0, 1)		uniform
Short activity lifespan	τ_1	($1/24$, 1)	day	uniform
Simulation time	T	1	year	fixed
Speed	μ	(0, $1/2$)		uniform
Transient time	T_0	2	year	fixed

Table 5.1. Parameters settings for global sensitivity analysis.

available on the website of the Wikimedia Foundation¹. Since rate parameters have a strong influence on the simulation time, ranges for these parameters have been chosen trying to strike a balance between exhaustiveness of the sensitivity analysis and wall clock time of the simulations.

A bit more problematic are the constant popularity term (c_p) and constant successes term (c_s) parameters. These are non-physical quantities and have never been studied before; the literature gives no guidance here. For both parameters we arbitrarily picked an interval that we felt to be large enough to be considered plausible.

Finally, the opinion dynamics parameters. It is clear that speed μ should not vary beyond $1/2$. Regarding the confidence ε , the literature on bounded confidence models in one dimension suggests that for $\varepsilon > 1/2$ the dynamics of consensus does not change noticeably. This should apply also to the dynamics of user participation in our model. We ran some simulation of the average activity lifespan that indeed confirm this. For the sake of further saving computational time we thus restricted the confidence parameter ε to the interval (0, $1/2$). Finally, the parameter of the rollback probability (r) was kept to range in its (0, 1) interval.

¹<http://stats.wikimedia.org>.

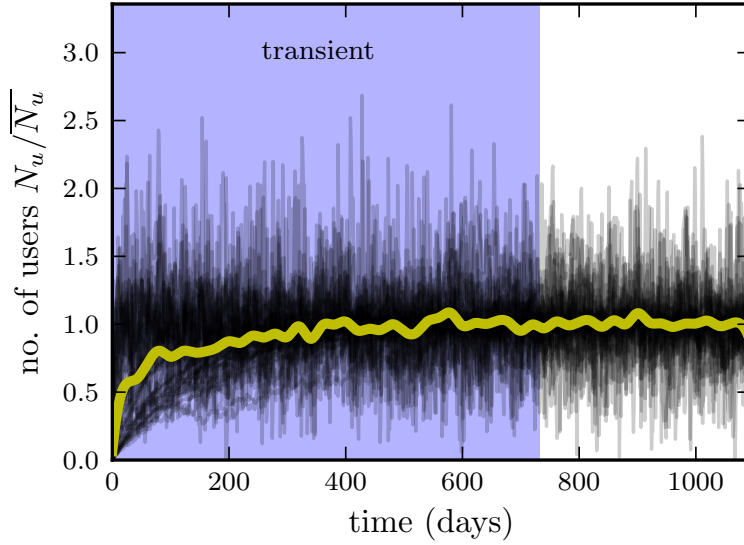


Figure 5.1. Transient time determination.

Transient

Due to its dynamical nature the initial number of users (N_u) requires a special procedure for determining its value. At each time step Δt a number of new users may join the community, while some of its current users may leave it. Thus at time t the number of active users $N_u(t)$ depends on both the rate of arrivals and the rate of departures. While the daily rate of arrivals is given by the parameter ρ_u , the rate of departures depends on the current size of the community $N_u(t)$ and, in general, on the dynamics of interactions between users and pages – which we cannot control directly. At equilibrium, however, these two rates are equal and the number of users is constant, that is, $N_u = \langle N_u(t) \rangle$, where the average $\langle \cdot \rangle$ is meant as a time average.

To determine N_u we run the model for a time equal to T_0 . The length of this transient phase was determined empirically. We plotted the daily number of users $N_u(d)$, $d = 1, 2, \dots$, for various values of the parameters of the model and we determine T_0 as the time after which all curves look stationary. Figure 5.1 reports the results of this exercise for a maximin Latin hypercube with 50 points. In the figure, the shaded region corresponds to the transient interval $(0, T_0)$. The value of T_0 is 730 days. Each curve is scaled by the average value \bar{N}_u computed over the interval $d \in [731, 1095]$. The red line is a B-spline fit of the average value of N_u across all curves, evaluated at 50 evenly spaced points along the x axis, and serves only as a guide to the eye.

During the transient phase we do not record any data, so that the estimation of τ , on which the sensitivity our analysis is based, does not reflect the dynamics of opinion formation during the transient.

In these simulations the initial number of pages N_p grows instead indefinitely. We do not allow pages to “leave” the community, although they might be selected less and less over time. The expected initial number of pages at the end of the transient is thus given by $\rho_p T_0$.

5.3.2 Factor screening via global sensitivity analysis

We sample a maximin Latin Hypercube design with 50 points using the intervals listed in Table 5.1. To sample a decent maximin design, we generate 10^4 hypercubes and select the one that maximizes (5.20). Our vector of parameters is thus:

$$\boldsymbol{\theta} = (\lambda_e, \rho_u, \rho_p, \varepsilon, \mu, c_s, c_p, r, \tau_0, \tau_1) \quad (5.22)$$

We compute the average activity lifespan $\langle \tau \rangle$ by taking the mean difference between the time of the first and last edits of each user. For each $\boldsymbol{\theta}$, we run 10 replications and average the values obtained. In case a replication does not produce any data point, for example when no agent produces more than one interaction, we take $\langle \tau \rangle = 0$. We have checked manually whether any value of $\boldsymbol{\theta}$ in the sample would produce no observation across all replications and found no problem of this kind.

Parameter	ρ	t	p -value	DOF
ε	0.85 ^{***}	9.99	0.0	39
c_p	-0.13	-0.81	0.42	39
c_s	0.042	0.27	0.79	39
λ_e	-0.16	-1.00	0.32	39
ρ_p	-0.36 ^{**}	-2.42	0.02	39
ρ_u	-0.28 [*]	-1.85	0.07	39
τ_0	0.54 ^{***}	4.06	0.0	39
r	-0.04	-0.22	0.82	39
τ_1	-0.002	-0.02	0.98	39
μ	-0.07	-0.42	0.67	39

Table 5.2. Partial correlation coefficients (^{***}: < 1%, ^{**}: < 5%, ^{*}: < 10%).

Table 5.2 reports on the partial correlation coefficients between $\langle \tau \rangle$ and any single parameter, controlling for the effect of all remaining parameters. We

can see that two parameters, ε and τ_0 , have a strong positive effect on average activity lifespan. Moreover, the correlations are statistically significant according to a t -test. The daily rate of page creations ρ_p has instead a negative effect, with $\rho = -0.36$, that is moderately significant. Other parameters attain little or no correlation with the activity lifespan, and a t -test cannot rule out the null hypothesis of truly uncorrelated variates.

We decided to investigate further the role of other parameters and see if a linear regression would give us more statistically significant results. Table 5.3 reports the results of this analysis. While for the confidence ε and for the long activity lifespan τ_0 the result of a strong positive effect is confirmed, for ρ_p the coefficient is almost null and not statistically significant. Moreover, the analysis shows also a weakly significant effect for the short activity lifespan τ_1 .

Parameter	Coefficient	Std. error	t	p -value
λ_e	0.022	0.071	0.31	0.75
ρ_u	0.07	0.078	0.86	0.39
ρ_p	-0.081	0.069	-1.19	0.24
ε	0.85***	0.068	12.47	0.0
μ	-0.1	0.073	-0.14	0.88
c_s	0.091	0.07	1.3	0.2
c_p	0.046	0.068	0.68	0.5
r	0.089	0.069	1.29	0.21
τ_1	0.13*	0.072	1.77	0.083
τ_0	0.46***	0.075	6.03	0.0

Table 5.3. Standardized linear regression. Coefficient of determination $R^2 = 0.83$, adjusted $R^2 = 0.79$.

We plot in Figure 5.2 scatters of the response versus input factors. Despite a good value of $R^2 = 0.83$ for the linear regression, it is obvious from the scatter plots that the response has a highly non-linear behavior, especially for the most important factor ε . In addition, we tried fitting a sigmoid function to τ as a function of ε but the results (not shown) suggested a poor-quality fit. This non-linearity indicates that both partial correlations and linear regression cannot tell us much about the actual share of the variance each parameter is responsible for.

We finally turn to the decomposition of variance. We fit a Gaussian process (GP) emulator average user lifespan data, obtained by running our simulator with the maximin Latin hypercube design. Roughly speaking, the meaning of using a GP emulator is the following: let us denote with $T = (\tau_1, \dots, \tau_{50})$ the observed – that is, simulated – activity lifespan responses. At an untried input

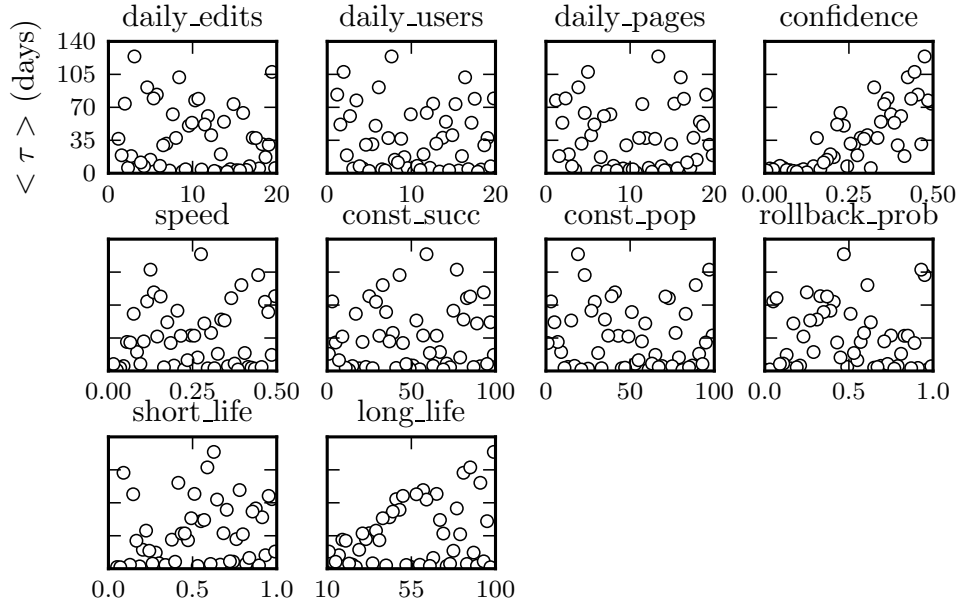


Figure 5.2. Scatter plots of $\langle \tau \rangle$ versus θ . Error bars (standard error over 10) are all smaller than the data points.

vector θ the GP approximation gives us a prediction of the average activity lifespan $\hat{\tau}(\theta) = E[\tau(\theta) | T]$. We can thus use $\hat{\tau}$ in lieu of our simulator to compute the sensitivity indices using the winding stairs methods.

We compute main and total interaction effect indices of each parameter using a ws matrix W with 10^4 rows. The results are shown in Table 5.4.

The total variance $\hat{\sigma}^2$ is also computed from W (remember that each column is a sample of independent observations). Given the uncertainty in the estimation of the total variance σ^2 , the fact that M_i is slightly negative for those parameters with $T_i \approx 0$ should not come as surprise. In fact, the way of computing the main effects does not guarantee that the estimates are always positive, so small negative values are not at all a surprise. It should be noted that the method for computing the indices does not even guarantee that the estimates satisfy $M_i < T_i$. Actually, Chan et al. [2000] showed that (5.16) and (5.17), used in conjunction with the Winding Stairs schema, tend to give better estimated of the total interaction indices than other methods, so we take the numbers shown in Table 5.4 as a sign of the goodness of the estimates.

In passing, it should be said that this is all credit to using a GP emulator in conjunction with the Winding Stairs method. For our 10 parameters, a ws matrix

Parameter	M_i	T_i
λ_e	-0.002	0.014
ρ_u	-0.003	0.02
ρ_p	0.003	0.027
ε	0.65	0.73
μ	-0.004	0.03
c_s	0.004	0.03
c_p	-0.005	0.016
r	-0.005	0.026
τ_1	0.002	0.03
τ_0	0.18	0.23

Table 5.4. Variance decomposition. Total variance : 635.365 days².

with 50 rows and with 10 replications for each input would require more than 5000 simulations. Computation time of a single model run varies considerably from a few seconds to several minutes, depending on the combination of input parameters for the editing rate λ_e , the rates of new users arrivals ρ_u , and the rate of new page creations ρ_p . In early simulations, for a sensitivity analysis with different parameters settings (most notably with values of τ_0 and τ_1 fixed to 100d and 1h, respectively) that we choose not to report on, the computation had taken several hours. With these settings, the same simulations took several days instead.

Using a GP approximation of τ allowed us instead to run a less expensive simulation over an LHS (for a reduced factor of 10 times less the number of runs), and to evaluate the ws matrix \mathbf{W} with many more rows than before, that is using a much larger sample size. Thus, being much less computationally expensive, the emulation approach allows us to produce much better estimates for the main and total interaction effect indices in a considerably smaller amount of time.

For the confidence ε and long term activity lifespan τ_0 parameters the decomposition of variance confirms the picture suggested by the two previous methods. This is further confirmed by Figure 5.3, in which we plot the main effect $Y(\theta_i)$ of each parameter θ_i as a function of the scaled parameter value, that is, ranging between 0 and 1 instead of between the endpoints listed in Table 5.1. To produce this plot we estimated integrals (5.11), for all 10 parameters, using the GP emulator.

From Table 5.4 we can see that both ρ_p and τ_1 , which were previously assigned a negative and a positive effect, account for almost a null fraction of the overall variance, and thus are not very much important. However, if we see

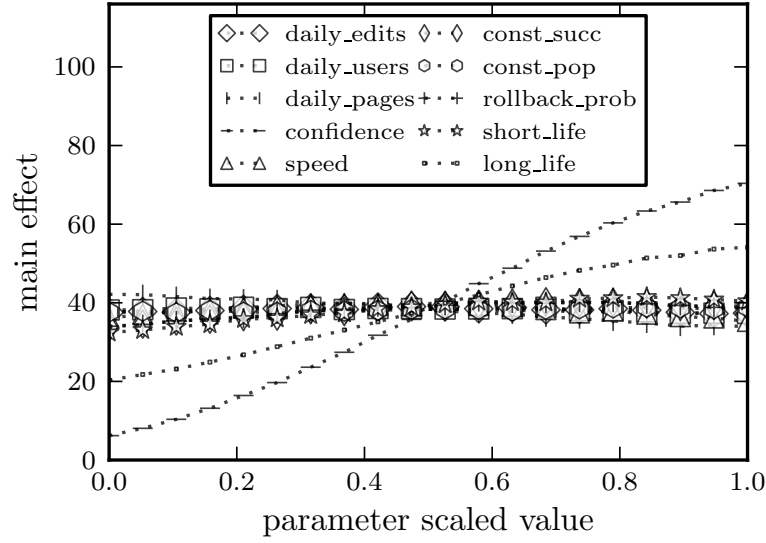


Figure 5.3. Main effects plot. The plot should be read in the following way: “high” values (> 0.8) of ε produce a value of the average activity lifespan > 60 days. Similarly for other parameters.

again figure 5.3, compared to all other parameters the main effect plot clearly shows a negative and a positive trend – except ε and τ_0 , of course. This nicely explains the results we got from the partial correlations and standardized regression for these two parameters.

It is worth noting that for ε the difference between T_i and M_i is 0.08 while for τ_0 is 0.05. Together, these two add up to almost three quarters (77%) of the total effects of all other parameters. While it is clear from the respective values of T_i that both ε and τ_0 have interactions with all other parameters, given this consideration it would be also interesting to understand how much the τ_0 and ε interact with each other, and what kind of interaction it is.

We can explore this question graphically looking at the so-called two-way interaction plot between ε and τ_0 . Given two parameters θ_i and θ_j , with $i \neq j$, we can compute $Y(\theta_i, \theta_j)$ evaluating a formula similar to (5.11), except that we now have a double integral instead of an integral in one variable; this time we simultaneously hold fixed two parameters instead of one. Thus this kind of plot is similar to the main effect plot, but produces a 3D response surface instead of a 2D curve. For the rest, estimation via a GP emulator is straightforward.

We produced 2-way interaction plots for the most important parameter ε and all other parameters – except c_s and c_p , which are scarcely interesting. The plots

are shown in figures 5.4-5.7. With the exception of τ_0 , for all other parameters we see only a very weak interaction with τ , for either low ($\varepsilon < 0.1$) or high ($\varepsilon > 0.4$) values of the confidence. For the rest, the average activity lifespan τ depends only on ε and is essentially independent from the other.

The pair $\{\varepsilon, \tau_0\}$ is the only exception to this. This confirms the intuition that τ_0 sets the support of the distribution of values for τ and that ε acts as a switch, controlling the transitions from a regime where a cluster of long-term users is able to emerge, to a regime where only short-term forms of participation are possible, due to low rate of successful edits in the peer-production process.

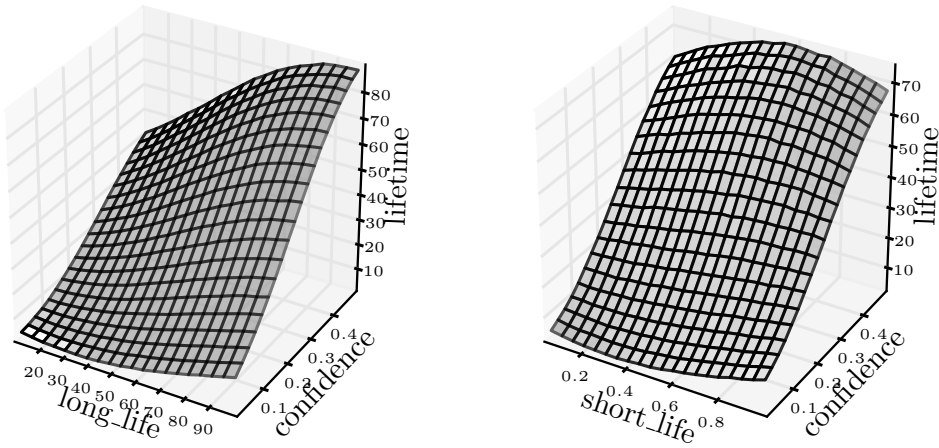


Figure 5.4. Left: long activity lifespan τ_0 versus confidence ε interaction effect plot. Right: short activity lifespan τ_1 versus confidence ε interaction effect plot.

5.4 Discussion

In this chapter we performed a factor screening for our peer production model via global sensitivity analysis. We tried three techniques, and we found that the one based on the decomposition of the output variance was the most suited for our case study, and the most informative.

We found that the confidence ε , the parameter that governs the update of opinions of both users and pages, accounts for almost all the variability in the response of the model. This may seem a bit surprising: at first glance, many

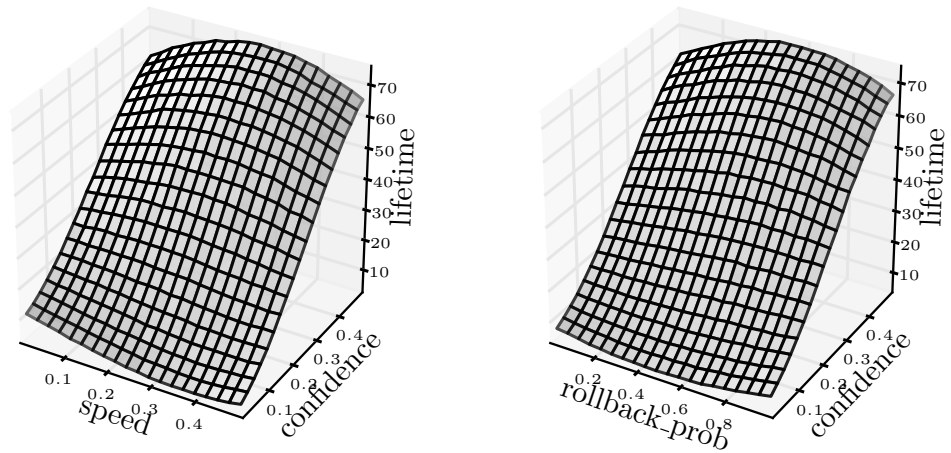


Figure 5.5. Left: speed μ versus confidence ε interaction effect plot. Right: rollback probability r versus confidence ε interaction effect plot.

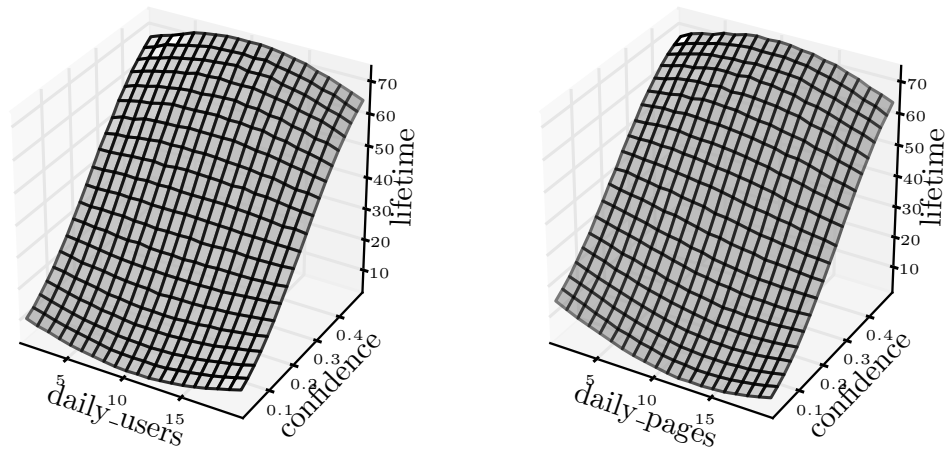


Figure 5.6. Right: daily rate of new users ρ_u versus confidence ε interaction effect plot. Left: daily rate of page creations ρ_p versus confidence ε interaction effect plot

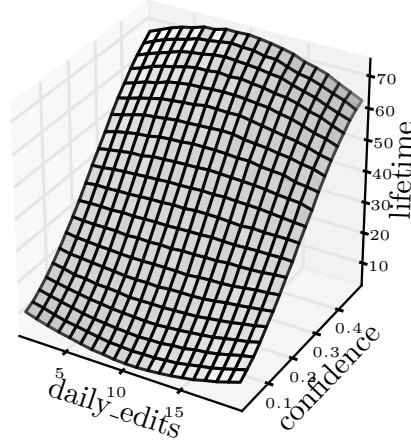


Figure 5.7. Daily rate of edits λ_u versus confidence ε interaction effect plot

other parameters – like the rate of edits λ_e , the rate of influx of users ρ_u , and the rate of influx of new pages ρ_p – may seem equally important.

It should be noted that the role of ε is of extreme importance in theoretical models of opinion dynamics under bounded confidence, like the one of Deffuant et al. [2001], since it governs the transition from a state of opinion polarization to one of full agreement. Our model of peer-production, then, preserves in spirit the main characteristics of those models. This result is somehow interesting, considering that models of opinion dynamics have found little application so far in empirical modeling.

Our use of global sensitivity analysis does not end with this chapter though. In the next chapter we will perform a model calibration using an indirect inference technique, and will apply again the machinery of variance decomposition to assess the predictive ability of the method.

Chapter 6

Simulation-based model calibration

6.1 Introduction

After having introduced our model of peer production and performed the factor screening exercise, we move to another part of the study of our model: the comparison of its predictions to empirical data from existing online communities, in our case the five versions of Wikipedia we already encountered in Chapter 3.

But first, it is worth briefly discussing what we mean by “comparison to empirical data”. A model, in the broadest sense possible, is just a simplified representation of an empirical phenomenon. Often, but not always, a model is equipped with a number of unknown parameters. By tweaking them appropriately, we can use it to show how simple stylized facts about the phenomenon under study can be reproduced by means of its (hopefully simple) rules. Agent-based modeling lends itself naturally to this “generative” approach (Bonabeau [2002]), which is especially fashionable in the Artificial Life community; simply put, it predicates that *“if you can’t grow it, you can’t explain”* (Epstein and Axtell [1996]). At this level the modeling exercise thus provides a useful way to filter, among all the possible mechanisms at work in the system under study, only those that are actually able to qualitatively reproduce its main phenomenological features.

However, if we want to understand a system at a quantitative level, we cannot merely show that, for a suitable range of values of the parameters of our model, we are able to reproduce the patterns in our data. We also need to see whether the values of these unknown parameters are compatible with the empirical data: it is thus the duty of the scientist to estimate the values of these parameters so that the output of the model resembles the empirical data on the phenomenon we set out to study and, since empirical data are usually affected

by noise and other forms of errors, quantify the uncertainty associated with these estimates.

In our case, the first questions we have to ask ourselves is how we should perform this empirical evaluation, that is, what statistical tools we can use, and what answers we can get from them. It is important to understand that the way we answer these questions will depend on the type of model we want to evaluate. Our case, in fact, does not yield easily to analytical investigation, so for example the classic machinery of statistics (e.g. maximum likelihood) is not available to us. Rather, we are in the (increasingly conventional) case of a computational model, that is, a model from which we can simulate. Given a value for each of the parameters, we can reproduce any of several different plausible scenarios of the evolution of a peer production community. We obviously run into the problem outlined above: what tells us that those are the *right* scenarios?

The factor screening told us merely that ε and τ_0 are two important parameters with respect to the average activity lifespan of users, in that they are responsible for much of the variation of these response variables. But it cannot tell us anything about what combination of values all parameters should take to best reflect the empirical data. Thus we have to resort to some form of inference. There are several statistical techniques to deal with parameter inference for computer codes. These techniques have been developed in number of contexts (engineering applications, population biology to cite a couple) where computational models are widely used, and yet quantitatively precise statements must be made about the predictions of a model.

The computational model fitting technique that we present here is inspired by the indirect inference methodology (Gouriéroux et al. [1993]). This technique predicates the use of an *auxiliary* model to match empirical data with synthetic simulations. Therefore, we will briefly describe it.

6.2 Background: Indirect inference

Indirect inference is a technique used to fit models to empirical data when maximum likelihood estimation is either unfeasible, or simply too complicated from a computational point of view. The maximum likelihood approach could be unfeasible because the model does not yield itself to writing down the likelihood function in analytic form. Many agent-based models fall into this category – and our case does too. But even if we had the likelihood function, sometimes it is just hard to optimize it. Those examples that motivated the development of indirect inference – dynamical models with latent variables – are often found in

Economics (cf. Gouriéroux et al. [1993]; Smith [2008]). In that context, one needs to compute an integral over all possible histories and values of the latent variables. There are approaches to doing this, like the Expectation Maximization (EM) algorithm (cf. Dempster et al. [1977]; Neal and Hinton [1998]) but they are often messy and complicated. The indirect inference method was introduced by Smith [1993] in the context of vector auto-regression models (VAR) in econometrics.

Indirect inference is an extension of the method of simulated moments (McFadden [1989]), and it requires only that we can simulate from our model and produce synthetic data sets from it. The idea of indirect inference is to introduce a second model, which in general is not the correct model for our phenomenon, but that easily fits the data. This model gives a criterion of agreement between the synthetic data and the observed ones, and therefore is called the auxiliary model. In order to ensure identifiability of the parameters, the only practical requirement we need from the auxiliary model is that it has at least as many parameters as we have in our simulation model. For example, if we were to fit a model that produces a time series, an option could be an auto-regressive model.

Under suitable conditions the indirect inference technique provides a consistent and asymptotically normal estimator (Gouriéroux and Monfort [1996]). It is commonly used in economics to validate agent-based models (Bianchi et al. [2007]), and in population biology (Ellner and Guckenheimer [2006]). A variant of indirect inference, used also in population biology, is called simulated quasi-maximum likelihood (Smith [1993]; Kendall et al. [2005]; Wood [2010]) when the likelihood function of the auxiliary model is used in lieu of that of the simulation model.

Let us consider a model \mathcal{M} with p unknown parameters $\theta = (\theta_1, \theta_2, \dots, \theta_p)$, and n independent, identically distributed observations from an empirical process $\mathbf{x} = (x_1, x_2, \dots, x_n)$, the normal MLE approach would go on maximizing the likelihood function:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \mathcal{L}(\mathbf{x}; \theta) \quad (6.1)$$

Here we assume that (6.1) is either intractable or that \mathcal{L} is unavailable in analytic form, which is the case for our peer production model.

The auxiliary model \mathcal{M}_a has instead a vector of parameters β . Let us denote by $\hat{\beta}$ the estimated vector of auxiliary parameters from the empirical data \mathbf{x} . This estimate may be obtained via maximum likelihood or any other suitable method. As we said, given a value of θ , we can simulate from our simulation and produce the synthetic dataset $\mathbf{x}^{(\theta)}$, where the superscript just reminds us

that the distribution of $\mathbf{x}^{(\theta)}$ will depend on θ . We can apply to $\mathbf{x}^{(\theta)}$ the same estimation technique we used for $\hat{\beta}$ and obtain $\hat{\beta}(\theta)$. At this point, we want to pick, as our estimate, a vector θ such that the difference between $\hat{\beta}$ and $\hat{\beta}(\theta)$ is as small as possible. Thus the indirect inference estimator of the true parameter vector θ_0 is obtained by minimization of a quadratic form:

$$\hat{\theta}_{\text{IND}} = \arg \min_{\theta} \left(\hat{\beta} - \hat{\beta}(\theta) \right)^T W \left(\hat{\beta} - \hat{\beta}(\theta) \right) \quad (6.2)$$

where W is a positive definite matrix that is used to give more or less weight to the auxiliary parameters, based on their sensitivity with respect to θ . In practice the identity matrix is often used. Equation (6.2) is not the only possibility; other metrics are available, see Smith [2008]. Moreover, if the asymptotic distribution of $\hat{\beta}$ is normal, a common trick to enhance convergence is to generate via simulation S different realizations of the data $\mathbf{x}_1^\theta, \dots, \mathbf{x}_S^\theta$ for a given θ , fit each of them to the auxiliary model, and then take the average.

The intuitive idea behind this is that, if the auxiliary model \mathcal{M}_a is able to capture the main feature of the data, that is, if it is sensitive enough to changes of θ , then it induces an invertible function $\beta(\theta)$ of the parameters of our model. Then solving equation (6.2) basically amounts to invert this function, so that we find the value of θ associated to the estimate $\hat{\beta}$. Under the assumption that the empirical data have been generated by a “true” value θ_0 , this amounts to finding its estimate under model \mathcal{M} .

The choice of a good auxiliary model \mathcal{M}_a is critical here. In the calibration of our peer production model we performed several diagnostic checks in order to ensure that the required condition on $\beta(\theta)$ is satisfied. But before getting into this, we should first describe in detail our calibration technique.

6.3 Indirect inference of unknown density

In this section we describe our custom inference technique. The goal of our estimation technique is to calibrate the peer production model using data on the lifespan of user participation from the existing Wikipedia communities we already saw in chapter 3. Since we know that user lifespan follows a bimodal distribution, a straightforward choice of the auxiliary model is a mixture of log-normal distributions. Thus the auxiliary parameters will be:

$$\beta = (\mu_1, \dots, \mu_k, \sigma_1, \dots, \sigma_k, \pi_1, \dots, \pi_{k-1}) \quad (6.3)$$

where k is the number of mixture components (see (3.6)). Estimates are ob-

tained using the usual Expectation Maximization algorithm (see chapter 3). We tested both the regular EM and the one for truncated data. Mixture models are an interesting option for indirect inference because we can increase the number of parameters by allowing the mixture to be composed of more components, even though too many components will eventually result in some auxiliary parameters being insensitive to changes in θ .

With a classic econometric model (such as those presented in Gouriéroux and Monfort [1996]) a straightforward way to implement (6.2) would be to use a double loop. The outer loop would consist of an optimization routine such as Newton's method, while the inner loop would, given a value of θ , simulate several synthetic samples and estimate $\hat{\beta}(\theta)$. This would then be used by the optimization routine to perform the search for the minimum of (6.2) in the space of the parameters of the simulation model.

The estimation technique we actually use differs from the classic indirect inference methodology presented above in a few details. This different approach is motivated by the peculiarities of our case study. The above scheme, in fact, poses the same technical difficulties we encountered during the factor screening exercise of Chapter 5: compared to the econometric models with latent variables that are usually fitted via indirect inference, our model is more demanding from the computational point of view, since simulations can take, depending on the value of the parameters ε and τ_0 , from several minutes up to hours to complete. Therefore, given the same problem, we resorted to an approach that, in spirit, resembles the one employed for the sensitivity analysis, namely we used a surrogate model instead of the actual peer production simulator.

Similar approaches are usually taken when the output of the model is either multivariate (and has many variables), or when it has a functional form, for example a time series. In both cases there are multiple variables, but in the former the dimensionality is fixed, whereas in the latter it is not. If these are the cases, classic dimensionality reduction techniques, for example Principal Component Analysis (PCA), are usually applied (cf. Dancik et al. [2010]). This time, however, we cannot use the Gaussian Process estimation technique directly on the output of our peer production model. The reasons for this are:

- The normality assumption does not hold – we know the data follow a bimodal distribution.
- The dimensionality of the output of the model (i.e. a full sample of lifespan observations, and not a scalar or multivariate output) does not allow the application of the GP technique directly.

To overcome these added limitations we decided to perform a preprocessing step on the output of the model: we first fit the lifespan data to a mixture model using the Expectation Maximization algorithm, and then we estimated the sufficient statistic of the mixture using Gaussian processes. In other words, since the output of the simulation is a sample from an unknown distribution, i.e. a density functional, instead of applying the Gaussian process directly to our model, we first use the mixture model as a clustering technique, and then apply the GP to provide a surrogate of $\hat{\beta}(\theta)$, that is, of the mapping between the parameters of the peer production model and the sufficient statistic of the mixture of log-normals.

Our calibration scheme is summarized in figure 6.1. In the figure, the gray circle in the top row corresponds to the agent-based simulation step, while blue rectangles to parameter estimation steps. The first step is to take an input sample from the space of the parameters of the computer model, $\theta_1, \dots, \theta_N$. We used again a space-filling design, generated by sampling a min-max Latin hypercube sample. For simplicity, in the diagram we show only one simulation per input site θ . In practice, for each θ_i , $i = 1, \dots, N$, we simulate from the peer production model multiple times (in particular, $R = 10$), use EM on each to obtain the parameters of the mixture, and then average. The result of the simulation step is the matrix B_θ , where each row is the result of the averaging over these R replications. We then use B_θ to approximate $\beta(\theta)$ with the GP emulator.

The rest of the procedure follows the classic indirect inference technique: we separately produce auxiliary estimates $\hat{\beta}$ from the empirical sample of user activity lifespan observations τ , and use this, together with $\hat{\beta}_{GP}$ to perform the minimization of (6.2). To perform the minimization we can use either the simplex algorithm (Nelder and Mead [1965]) or the bound constrained optimization algorithm BFGS (Byrd et al. [1995]). Both are conveniently implemented in the open source SciPy scientific library.

Not all parameters are inferred with the above technique. In the next section we describe in detail the full setup we used to perform simulations for the calibration of our model.

6.4 Simulation design

The indirect inference technique requires us to perform simulations from the model according to an experimental design: in our case, a randomized block design based on minimax Latin Hypercubes. With the exception of the parameters on which we are going to perform the indirect inference, all other input

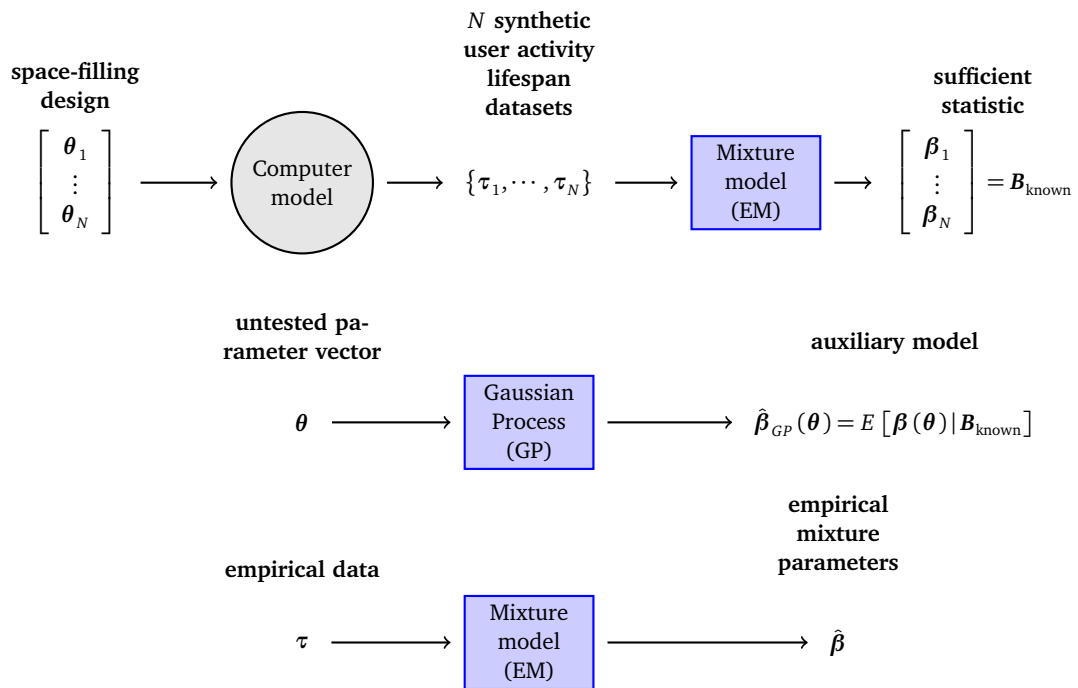


Figure 6.1. Indirect inference model calibration

variables of the model must be set to some value that allows the response of the model to be compared with the empirical data in the best possible way.

In general, among the parameters of our peer production model there are both process variables, such as the rate ρ_u at which new users enter the system, or the confidence parameter ε , and control variables, such as the length of the simulation interval T . For the latter type, there are usually obvious choices of their values, e.g. $T = t_0 - t_1$ where t_0 corresponds to the date of creation of the project, and t_1 to the last time stamp recorded in the revision metadata, for example.

All other process variables that are not going to be the object of the indirect inference calibration must be estimated in other ways. In practice, we perform the indirect inference on all those process variables that we cannot readily estimate from the data. There are:

- Confidence ε .
- Speed μ .
- Initial number of successes c_s .
- Initial popularity c_p .
- Rollback probability r .¹

This leaves us with the following parameters that can be instead directly estimated from data.

- Daily rate at which new users join the community ρ_u .
- Daily rate at which new pages are created ρ_p .
- Average user activity lifespan (long-term) τ_1 .
- Average user activity lifespan (short-term) τ_0 .

¹In theory, we could estimate also this quantity from the data. One crude way would be to inspect the comment that is present among the revision metadata. The Mediawiki software, in fact, automatically adds a template text to the revision comment, whenever a user performs a rollback. Another, more sophisticated approach, would be to analyze the text of each revision and see if it matches other previous revisions. This can be done efficiently using MD5 hashing. Both approaches, however, are quite time-consuming, and thus we decided to include this parameter in the group estimated via indirect inference.

Finally, there is a group of parameters that we estimated by trial and error. These process variables were too difficult to estimate from the empirical data themselves but, on the other hand, could be set to reasonable defaults, given their straightforward physical interpretation. These are:

- Daily rate of editing sessions of a user λ_a .
- Average number N_a of session edits after the first one.
- Homogeneous rate λ_e at which additional edits are performed during a session.

The list of parameters included in the block design is given in table 6.1. These specifications are the same across all simulations, i.e. all datasets have been fitted using the same block design specifications (but different designs have been sampled each time).

Parameter	Variable name	Symbol	Min	Max
Rollback probability	rollback_prob	r	0	1
Speed	speed	μ	0	$1/2$
Confidence	confidence	ε	0	$1/2$
Const. successes	const_succ	c_s	0	100
Const. popularity	const_pop	c_p	0	100

Table 6.1. Parameters to be calibrated via Indirect Inference. Block design specification for all datasets.

The values of the other parameters, that is, those we fit directly from data, is given, one for each dataset, in table 6.2. In the next subsections we cover the estimation of such parameters (ρ_u , ρ_p , τ_0 , and τ_1) in detail.

6.4.1 Time scales of user lifespan

Perhaps the most important parameters we fit separately are the two time scales τ_0 and τ_1 . These are used to compute the activity lifespan of any user during the simulations and, as we saw from the factor screening, have an important impact on the overall lifespan statistics.

The approach we took was to estimate these two values directly from the data we wished to fit via indirect inference. Ideally, given a clustering of the user in two classes, the short-term users and the long-term users, both τ_0 and τ_1 should be computable from the observed user lifespans τ and the information

Parameter	Code name	Symbol	Value
All			
Daily sessions	daily_sessions	λ_a	1
Hourly edits	hourly_edits	λ_e	1 min^{-1}
Session edits	session_edits	N_a	2
Portuguese			
Daily rate of users	daily_users	ρ_u	9.40 d^{-1}
Daily rate of pages	daily_pages	ρ_p	6.55 d^{-1}
Short activity lifespan	short_life	τ_1	17.12 min
Long activity lifespan	long_life	τ_0	$1.01 \times 10^3 \text{ d}$
Simulation time		T	$3.03 \times 10^3 \text{ d}$
Italian			
Daily rate of users	daily_users	ρ_u	4.06 d^{-1}
Daily rate of pages	daily_pages	ρ_p	$6.4 \times 10^2 \text{ d}^{-1}$
Short activity lifespan	short_life	τ_1	15.43 min
Long activity lifespan	long_life	τ_0	$1.35 \times 10^3 \text{ d}$
Simulation time		T	$2.96 \times 10^3 \text{ d}$
French			
Daily rate of users	daily_users	ρ_u	2.19 d^{-1}
Daily rate of pages	daily_pages	ρ_p	$1.11 \times 10^3 \text{ d}^{-1}$
Short activity lifespan	short_life	τ_1	15.81 min
Long activity lifespan	long_life	τ_0	$1.32 \times 10^3 \text{ d}$
Simulation time		T	$3.04 \times 10^3 \text{ d}$
German			
Daily rate of users	daily_users	ρ_u	13.12 d^{-1}
Daily rate of pages	daily_pages	ρ_p	$8.9 \times 10^2 \text{ d}^{-1}$
Short activity lifespan	short_life	τ_1	15.85 min
Long activity lifespan	long_life	τ_0	$1.58 \times 10^3 \text{ d}$
Simulation time		T	$3.09 \times 10^3 \text{ d}$
English			
Daily rate of users	daily_users	ρ_u	$9.57 \times 10^2 \text{ d}^{-1}$
Daily rate of pages	daily_pages	ρ_p	$5.70 \times 10^3 \text{ d}^{-1}$
Short activity lifespan	short_life	τ_1	20.35 min
Long activity lifespan	long_life	τ_0	$1.31 \times 10^3 \text{ d}$
Simulation time		T	$3.09 \times 10^3 \text{ d}$

Table 6.2. Experimental design for calibration.

of group membership given by the latent variable computed by EM – the so-called “responsibilities”. In practice, instead of running EM we can take a much simpler approach, and just define a hard threshold τ_t for the lifespan of a user: observations of τ that are less than τ_t are assigned to the cluster of short-lived users, while observations of $\tau > \tau_t$ to the long-lived one.

Once we have performed this (rather crude) form of hard clustering, we compute suitable descriptive statistics that we take as the estimates for our two parameters, that is, $\hat{\tau}_0$ and $\hat{\tau}_1$. We tested several values of τ_t , and eventually settled for $\tau_t = 3\text{h}$ (fig. 6.3). As for the statistics we used, we computed both median and mean activity lifespan of each log-normal component. Our objective was to match the user activity lifespans produced by our model, therefore we performed some simulations with both values, adjusting the value of ε by hand, and found that the mean provided a more reasonable estimate (see fig. 6.2). Larger values of the threshold (in fig. 6.4, $\tau_t = 7\text{d}$ is reported) produce poor-quality estimates both for the mean and the median.

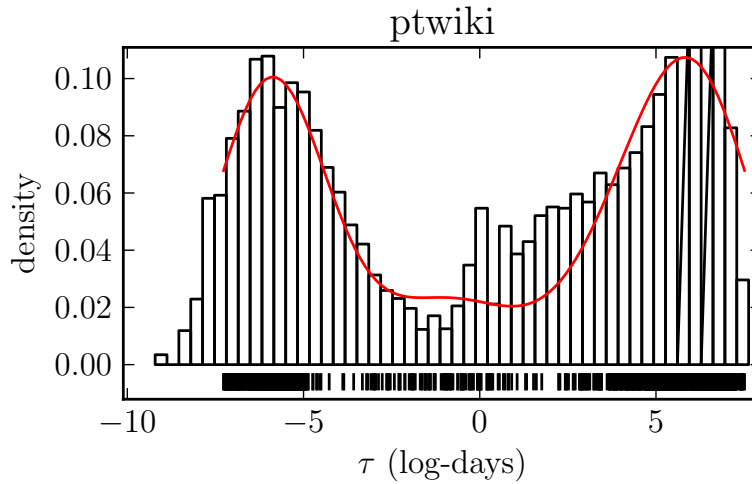


Figure 6.2. Test simulation for the estimation of the user lifespan scales $\tau_{0,1}$. Histograms: empirical data from the Portuguese Wikipedia. Black bottom vertical lines: simulated observations. Red line: nonparametric (i.e. kernel) density estimate of simulated data. Simulation was performed with mean of clustered data and $\tau_t = 3\text{h}$. Confidence $\varepsilon = 0.24$. Other parameters estimated from data of the Portuguese Wikipedia.

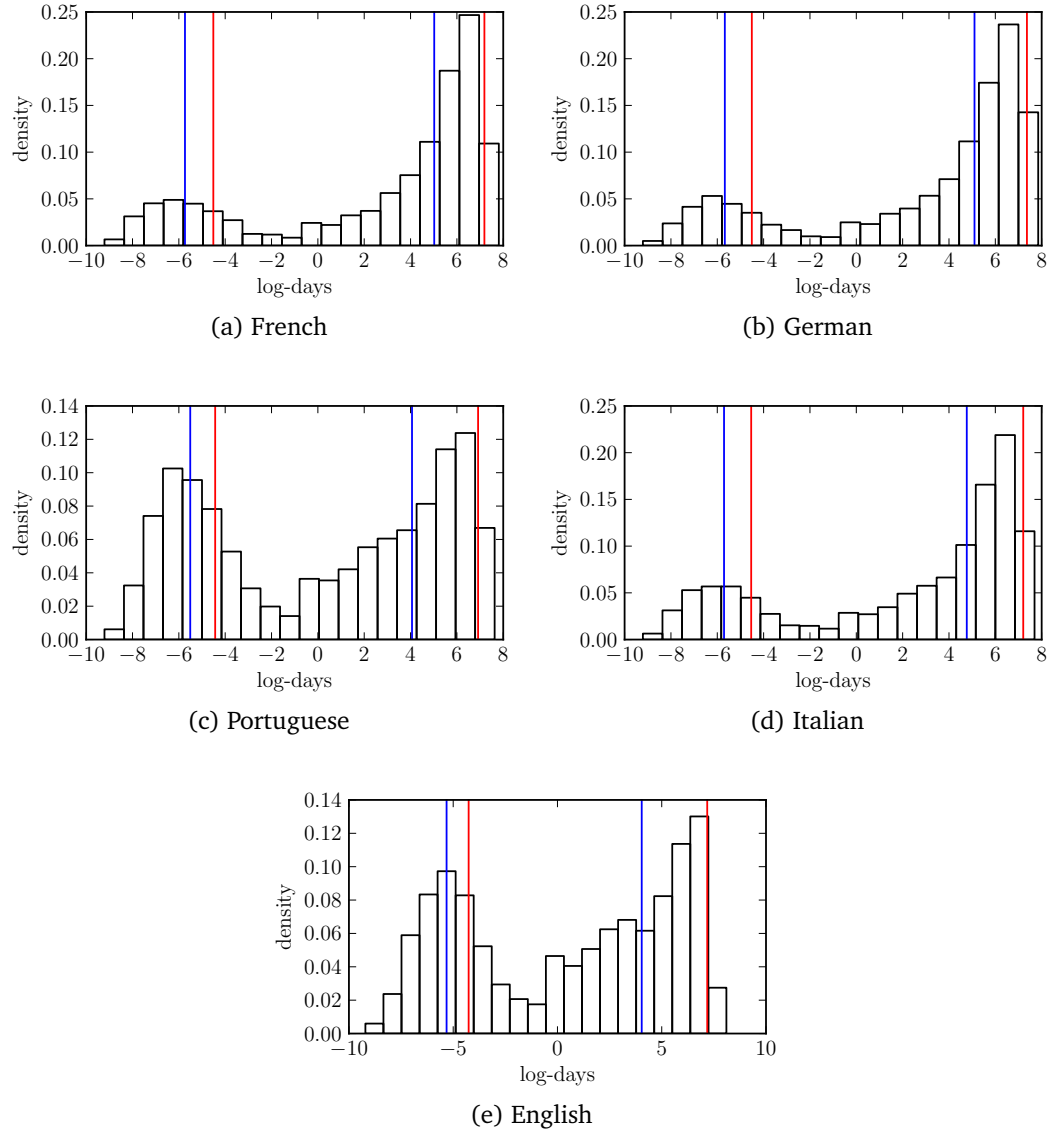


Figure 6.3. Mean (red) and median (blue) user lifespan. Empirical datasets. Clusters separated at $\tau_t = 3h$.

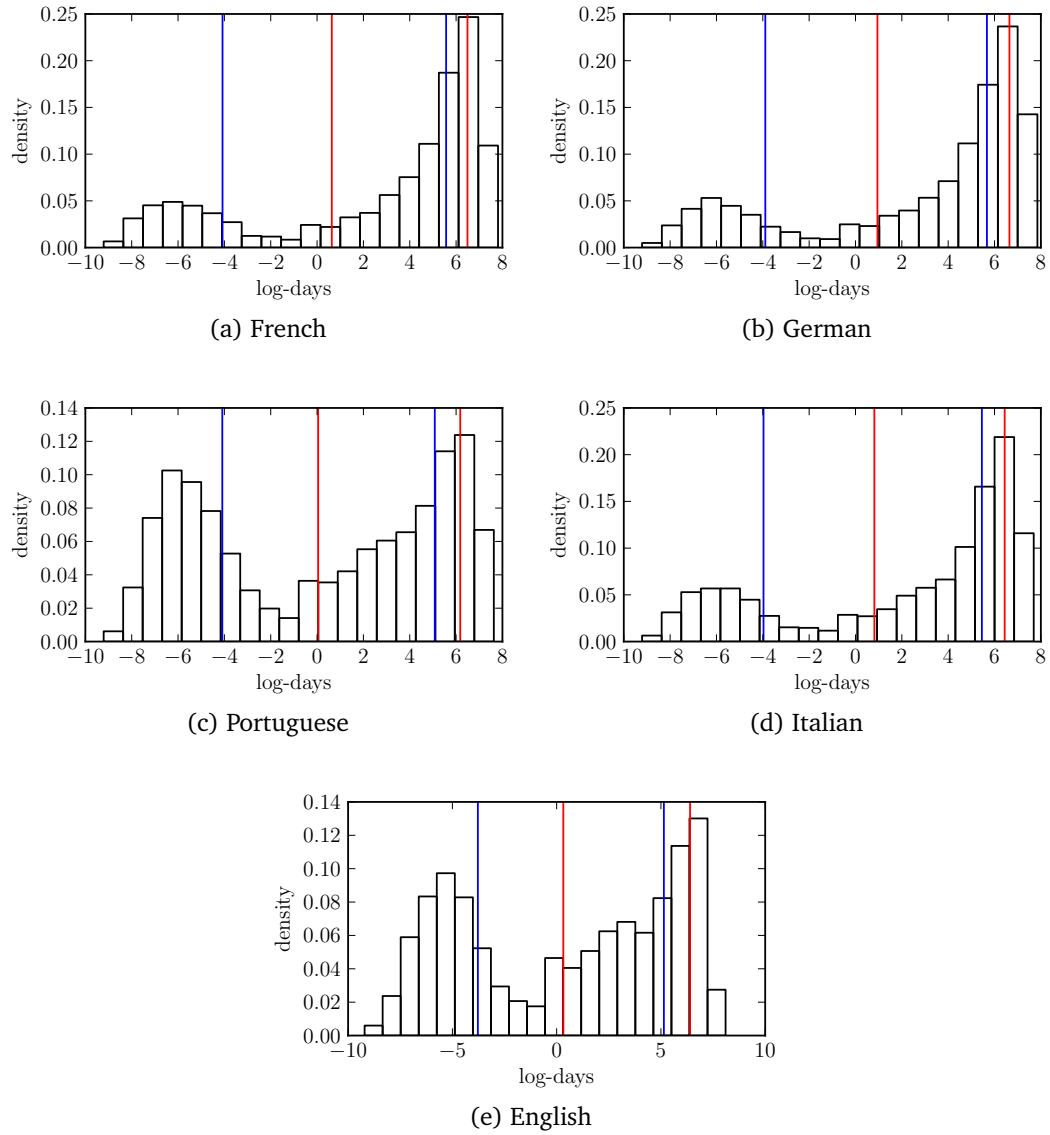


Figure 6.4. Mean (red) and median (blue) user lifespan. Empirical datasets. Clusters separated at $\tau_t = 7d$.

6.4.2 Activity rates

Here we want to quantify the rates of activity for two processes: the arrival of new users, and the creation of new pages. In our model these two processes are Poisson processes with a homogeneous rate of activity. In chapter 7 we will look at the problem of quantifying user activity in a more realistic way. Thus, our objective here is to quantify the average activity rate for both processes, so that we set the right scale of both processes for our calibration simulations. We estimate both rates from data.

Rate of new users

To compute the rate of new users joining the community, we need to know the day each user joined the community. This datum is usually recorded, as a full timestamp, whenever a user registers his or her account. In the Wikipedia database, this piece of information is recorded in the “user” table, alongside all other user details, such as password, email, and so on. Given the sensitivity of this information, the Wikimedia Foundation does not include this table in the snapshots dumps of the Wikipedia database.

Therefore, we resorted to using the day of the first edit as a proxy for the time of registration. To compute the rate at which new users enter the system, we thus just group by day (i.e. calendar date) and count how many users made their first edit on each specific day. Then, we take the average rate as the estimate for the parameter of our model: $\hat{\rho}_u = \overline{\rho}$.

Rate of new pages

We treat page creations as an independent process from user’s editing activity. To compute $\hat{\rho}_p$, we use a similar method to the one for the rate of new users. The time of creation of any page is present in the data, being namely the time of the first contribution. We then group by calendar date and again take the average.

6.5 Simulation and Diagnostics

For each community a min-max Latin hypercube with 32 sites was sampled out of 10^4 random extractions. For the Italian, French, and Portuguese communities simulations of each site of the hypercube were repeated 10 times. For the German 5 times. Each site was fitted to either a truncated or non-truncated

Gaussian mixture model (GMM). The number of components of the mixture was set to $k = 2$ and $k = 3$. The sufficient statistic of the mixture model was computed for each realization and then the results were averaged over all repetition of the same input site.

Simulations for the English data were attempted but could not be completed. The reason for this lies in the scale of activity of the system. The attentive reader will have noted that the rate of new users joining the system ρ_u is almost two orders of magnitude larger than for the German Wikipedia – the second largest community in our dataset. The rate ρ_p at which new pages are created presents similarly out-of-scale values.

Since, for values of ε close to $1/2$ the average lifespan tends to τ_0 – whose magnitude is comparable to T – this means that the rate of events to be simulated grows indefinitely during the whole simulation interval $[0, T)$, making impractical to simulate at sites from the corresponding region of the parameter space.

Even though we deemed it enough to have successfully simulated 4 systems out of 5, we should nonetheless note that the problem of simulating systems with massive rates of events can be in principle mitigated by dropping the requirement of perfect sampling in the simulation algorithm (Gillespie [1977]).²

Thus, in the remainder of the chapter we will present the results of the calibration of the peer production model only for the data from the following localized communities: German, Italian, French, and Portuguese.

6.5.1 Approximation of auxiliary parameters via Gaussian Processes

In order to assess the quality of the calibration method we computed several diagnostic indicators. The most crucial component in the indirect inference technique is the auxiliary model. The choice of a good intermediate model is important because $\beta(\theta)$ must be able to capture the features of the data well enough to be able to discriminate between different choices of the model's parameters θ .

Identifiability may be hindered if multiple values of θ result in similar values of β . A quick way to check this is to plot β , parametrized by θ , and see if the curve crosses over itself at one or more points.

²Practically, instead of updating the global activity rates after each event, we would perform it at regular intervals. The assumption is that the events happening during a small time interval Δt are independent, and thus do not affect the global activity rates. Therefore Δt should be chosen not to be too large.

We allowed ε to range in the interval $(0, 1)$, and used the original (i.e. without edit cascades) peer production model to produce plots of $\boldsymbol{\beta}(\varepsilon)$. The auxiliary model here is a simple GMM with two components. This model has five parameters: two means (μ_1 and μ_2), two variances (σ_1 and σ_2), and one weighting coefficient (π_1). Therefore we need to project $\boldsymbol{\beta}(\varepsilon)$ in a lower dimensional space. Figure 6.5 reports few selected combinations of the all the ten possible pairwise choices of these parameters. The first two plots (μ_1 vs μ_2 and σ_1 vs σ_2) are parameterized implicitly by ε , while the last one reports the behavior of π_1 versus ε explicitly.

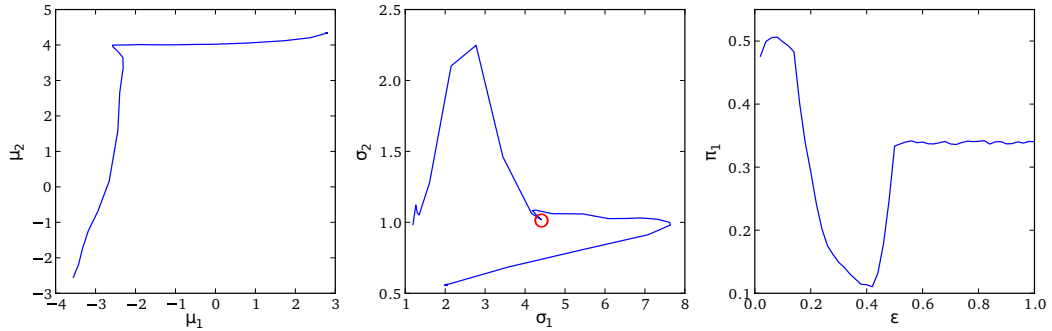


Figure 6.5. Auxiliary parameters as a function of ε . Auxiliary model: GMM with 2 components. Implicit and explicit parametric plots of parameter ε . Left: μ_1 versus μ_2 ; center: σ_1 versus σ_2 ; and right: π_1 versus ε .

The center plot shows a small kink (red circle), so we might expect some minor problem if the true ε is close to those values. Of course this visualization is incomplete, since we do not show all possible combinations between the components of vector $\boldsymbol{\beta}$. Moreover, the results shown in fig. 6.5 are only quantitatively different from those one would get using the actual model with edit cascades that we are actually going to calibrate; nonetheless they give an idea of what $\boldsymbol{\beta}(\boldsymbol{\theta})$ looks like.

Figure 6.6 instead shows the result of the GP approximation of the same auxiliary parameters (i.e. the sufficient statistic of the mixture model). Since the correspondence between model and auxiliary parameters $\boldsymbol{\theta} \mapsto \boldsymbol{\beta}$ is a many to many mapping, we fit as many Gaussian processes (i.e. many-to-one mappings) as auxiliary parameters, in a manner similar to the sensitivity analysis of chapter 5. The blue band represents the 95% confidence interval. In general we see that the approximation is very tight already with as few as 50 observations from the real response surface of the auxiliary model, i.e. $\boldsymbol{\beta}(\varepsilon)$.

We can immediately notice that for $\varepsilon > 0.5$ the auxiliary parameters record

no change. This is obvious, since for those values of the confidence parameter the dynamics of agreement always result in a full consensus case – and thus the average lifespan of the population is τ_0 . This observation actually motivated our choice to restrict the range of ε to the interval $(0, 1/2)$ in the calibration simulations.

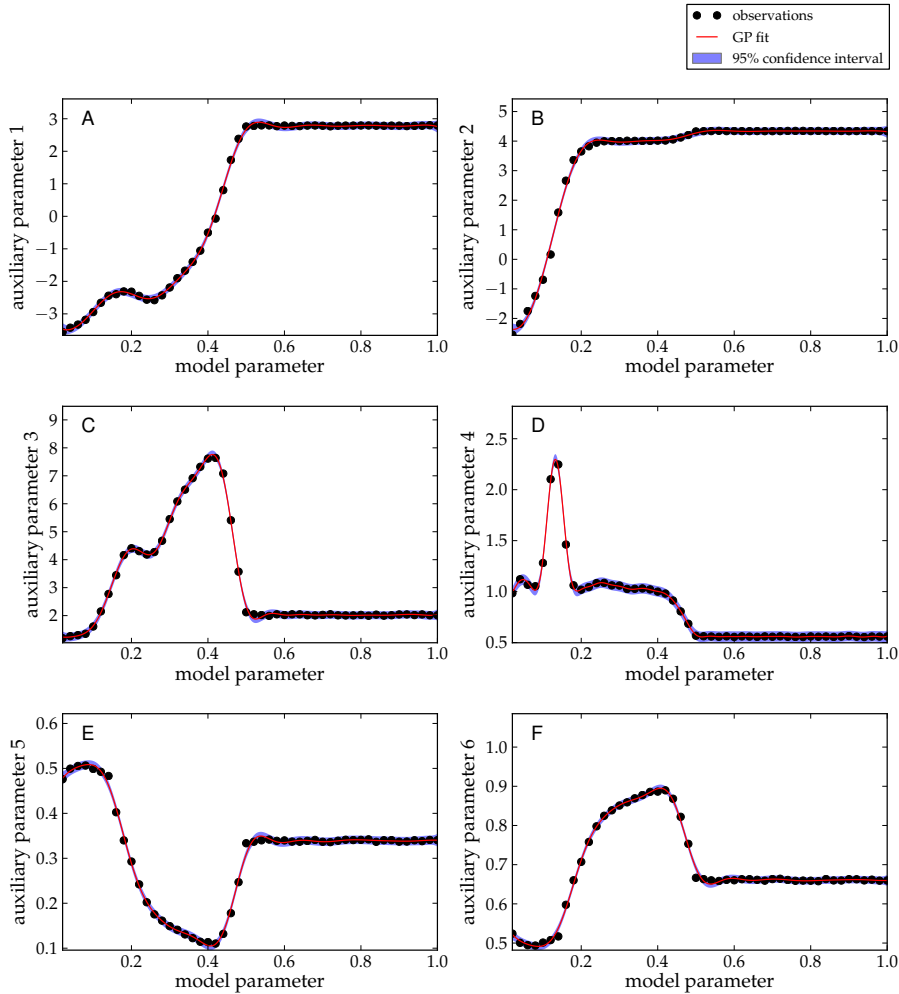


Figure 6.6. Emulation via Gaussian Processes of GMM sufficient statistic (Auxiliary parameters 1–6). A: μ_1 ; B: μ_2 ; C: σ_1 ; D: σ_2 ; E: π_1 ; F: π_2 . Each parameter is fitted to a univariate GP. Model parameter: confidence ε .

6.5.2 Sensitivity analysis of auxiliary parameters

Looking at the response surface gives us some intuition about the behavior of the auxiliary model but we would also like to quantify how sensitive each auxiliary parameter is to changes in the inputs of our model. Moreover, we can choose the number of mixture components to fit, and also whether to use the truncated mixture model or not. Thus, we would like to know what the best auxiliary model is among all possible choices we have.

Plotting $\beta(\theta)$ does not help much. The above questions can be answered by resorting once again to sensitivity analysis. This approach is useful because we can also use the sensitivity indices to define the matrix W of (6.2).

The results of the sensitivity analysis and of the whole calibration are shown in the following pages for the Portuguese (tables 6.3 and 6.4), Italian (tables 6.5 and 6.6), French (tables 6.7 and 6.8), and German (tables 6.9 and 6.10).³

Each auxiliary parameter is considered a response variable, of which we can compute a variance. The parameters of the peer production model are the “input” factors of the sensitivity analysis. For them we compute the main effect and the total interaction effect (see chapter 5 for the respective definitions.) The auxiliary parameters with the highest variance are the locations μ of the mixture components (i.e. the means) and, to some extent, the variances σ . Among the means, μ_1 is usually the one that attains the highest variability: when $k = 2$ this is the location of the high lifespan component, while for a mixture with 3 components it is the intermediate one.⁴

From the decomposition of variance we can notice that, in general, the confidence parameter (ϵ) is responsible for most of the variability of the auxiliary parameters. This is true both individually (high main effect indices M_ϵ) and through interaction with other parameters (total interaction effect T_ϵ). However, across languages the numbers for M_ϵ and T_ϵ differ: for Portuguese and $k = 2$, with the only exception of variance-1 (i.e. σ_1^2), all values of M are greater than 0.5, while for Italian, French, and German this is not the case and, in particular, the main effect of mean-1 is much larger than that for mean-0. Similarly, for

³A note on terminology: since our datasets refer to different localizations, we should refer to e.g. “the data for the Portuguese-speaking Wikipedia community”, and not just to “the Portuguese Wikipedia”. We stick to this erroneous terminology just for the sake of readability, but the reader should always keep in mind this important distinction.

⁴Note that labeling of the mixture parameters does not match the labeling of the time scale parameters in the model, e.g. μ_0 is the mean of the short-term component, while the short-term lifespan parameter in the peer production model is τ_1 . This discrepancy is due to the way parameters are identified from the results of the EM algorithms. This discrepancy is somehow unavoidable since, in general, the number of mixture components needs not match the number of time scales in the peer production model.

$k = 3$, it is difficult to find a pattern for the M_ε s of different languages.

Truncation seem to decrease the difference in variability between the mixture means. When the truncated model is considered, mean-0 increases over mean-1 for $k = 2$, and respectively mean-1 decreases with respect to mean-0 and mean-2 for $k = 3$. Besides this, when main and total interaction effect indices are considered, there is no qualitative difference between a truncated model and the standard one.

To sum up, if we focus on the auxiliary parameters with highest variability, the sensitivity analysis tells us that these are most sensitive to changes in the confidence alone. High values of the ratio $M_\varepsilon/T_\varepsilon$ (e.g. for the German with a GMM with $k = 2$ components, $M_\varepsilon/T_\varepsilon \approx 82\%$) confirm that most of the effects of these parameters are in fact due to first order interactions, that is, changes to itself in isolation. However, we cannot exclude complex interactions between the parameters since, in general, $T_\varepsilon - M_\varepsilon$, i.e. the fraction of variance accounted by ε in conjunction with other parameters, is comparable to each $T_p - M_p$, $p \in \{r, \mu, c_s, c_p\}$.

6.5.3 Cross-validation

The sensitivity analysis gave us essentially a picture of what we should expect from the calibration of the model, but it cannot quantify how reliable our indirect inference technique is. This question can be answered by means of a cross-validation technique. Assuming that the empirical data have been generated by our peer production model from a true, unknown vector of parameter θ_0 , we ask ourselves: how good is the estimate $\hat{\theta}_0$ obtained by indirect inference?

Of course we do not know the true θ_0 (that's why we are calibrating the model after all!), so we cannot compute the distribution of the estimator. But we can simulate from our model and test the ability of the indirect inference technique to reconstruct the “true” θ_0 that generated our synthetic data. Here we use a classic leave-one-out technique to perform this: using the simulated datasets evaluated on the minimax hypercube, we consider each pair $(\theta, \{\tau_i\}_{i=1}^{N_\theta})$ at a time, where τ_i is a lifespan observation and N_θ is the number of observations in the sample associated to θ . Given such a test pair, we put it away, which leaves us with a Latin hypercube with $N - 1$ sites. We then perform the Gaussian process approximation of the auxiliary parameters on this reduced input sample, and use it to estimate $\hat{\theta}$, i.e. the vector of parameters that generated testing data $\{\tau_i\}_{i=1}^{N_\theta}$.

We can plot the observed $\hat{\theta}$ as a function of the expected θ , which should be a straight line with slope equal to one and intercepting zero. We performed

variable	variance	rollback_prob r	speed μ	confidence ε	const_succ c_s	const_pop c_p
GMM, $k = 2$						
Main Effects						
mean-0	2.456	0.033	0.071	0.533	0.044	0.004
mean-1	10.137	0.011	0.004	0.645	0.027	0.000
variance-0	1.158	0.013	0.024	0.637	0.124	0.007
variance-1	0.077	0.040	0.115	0.066	0.086	0.144
weight-0	0.050	0.005	0.006	0.873	0.024	0.001
Interaction Effects						
mean-0	2.456	0.161	0.241	0.706	0.186	0.057
mean-1	10.137	0.155	0.177	0.785	0.131	0.102
variance-0	1.158	0.144	0.092	0.755	0.174	0.045
variance-1	0.077	0.329	0.371	0.328	0.280	0.390
weight-0	0.050	0.017	0.032	0.936	0.099	0.015
GMM, $k = 3$						
Main Effects						
mean-0	1.503	0.023	0.038	0.389	0.095	0.020
mean-1	16.618	0.036	0.012	0.709	0.022	0.018
mean-2	11.331	0.031	0.028	0.698	0.055	0.027
variance-0	0.970	0.014	0.029	0.369	0.153	0.026
variance-1	1.288	0.016	0.044	0.309	0.125	0.068
variance-2	0.071	0.041	0.025	0.499	0.055	0.018
weight-0	0.015	0.118	0.012	0.145	0.050	0.089
weight-1	0.039	0.054	0.014	0.304	0.232	0.038
weight-2	0.031	0.024	0.039	0.500	0.197	0.010
Interaction Effects						
mean-0	1.503	0.148	0.227	0.608	0.378	0.096
mean-1	16.618	0.115	0.084	0.845	0.138	0.078
mean-2	11.331	0.124	0.150	0.780	0.116	0.086
variance-0	0.970	0.151	0.186	0.585	0.399	0.109
variance-1	1.288	0.264	0.337	0.465	0.272	0.306
variance-2	0.071	0.168	0.157	0.733	0.241	0.122
weight-0	0.015	0.290	0.259	0.506	0.288	0.382
weight-1	0.039	0.202	0.228	0.492	0.363	0.236
weight-2	0.031	0.103	0.228	0.560	0.324	0.100

Table 6.3. Decomposition of variance, GMM. Portuguese Wikipedia.

variable	variance	rollback_prob r	speed μ	confidence ε	const_succ c_s	const_pop c_p
Truncated GMM, $k = 2$						
Main Effects						
mean-0	2.758	0.035	0.065	0.678	0.030	0.010
mean-1	10.022	0.021	-0.007	0.660	0.056	0.027
variance-0	1.010	0.015	0.002	0.688	0.126	0.021
variance-1	0.074	0.030	0.095	0.097	0.083	0.111
weight-0	0.047	0.015	0.010	0.869	0.051	0.020
Interaction Effects						
mean-0	2.758	0.139	0.138	0.796	0.103	0.052
mean-1	10.022	0.160	0.172	0.769	0.137	0.101
variance-0	1.010	0.108	0.086	0.763	0.172	0.049
variance-1	0.074	0.340	0.361	0.343	0.310	0.360
weight-0	0.047	0.018	0.033	0.920	0.104	0.015
Truncated GMM, $k = 3$						
Main Effects						
mean-0	1.319	0.048	0.004	0.489	0.057	0.039
mean-1	12.702	0.017	0.003	0.841	0.008	0.012
mean-2	10.625	0.014	0.025	0.724	0.011	-0.001
variance-0	0.526	0.053	-0.005	0.587	0.053	0.027
variance-1	0.657	0.043	0.027	0.498	-0.001	0.015
variance-2	0.055	0.003	0.020	0.730	0.054	0.027
weight-0	0.027	0.002	0.012	0.901	0.011	-0.008
weight-1	0.008	0.054	0.173	0.178	0.107	0.069
weight-2	0.027	0.023	0.025	0.589	0.065	0.004
Interaction Effects						
mean-0	1.319	0.158	0.287	0.704	0.259	0.048
mean-1	12.702	0.070	0.044	0.916	0.075	0.046
mean-2	10.625	0.114	0.138	0.813	0.096	0.077
variance-0	0.526	0.161	0.209	0.771	0.204	0.027
variance-1	0.657	0.311	0.268	0.576	0.102	0.180
variance-2	0.055	0.098	0.087	0.811	0.131	0.072
weight-0	0.027	0.027	0.053	0.923	0.057	0.028
weight-1	0.008	0.262	0.326	0.389	0.326	0.235
weight-2	0.027	0.111	0.162	0.695	0.233	0.089

Table 6.4. Decomposition of variance, truncated GMM. Portuguese Wikipedia.

variable	variance	rollback_prob r	speed μ	confidence ε	const_succ c_s	const_pop c_p
GMM, $k = 2$						
Main Effects						
mean-0	0.778	0.073	0.072	0.230	0.078	0.045
mean-1	2.845	0.000	-0.005	0.768	0.006	-0.009
variance-0	0.581	0.073	0.072	0.240	0.088	0.047
variance-1	0.020	0.006	0.072	0.264	0.135	0.052
weight-0	0.043	-0.024	-0.022	0.800	0.097	-0.013
Interaction Effects						
mean-0	0.778	0.223	0.389	0.504	0.365	0.085
mean-1	2.845	0.104	0.048	0.902	0.112	0.055
variance-0	0.581	0.222	0.356	0.502	0.370	0.085
variance-1	0.020	0.157	0.346	0.491	0.427	0.150
weight-0	0.043	0.031	0.056	0.833	0.160	0.036
GMM, $k = 3$						
Main Effects						
mean-0	0.181	-0.001	0.038	0.099	0.109	0.093
mean-1	11.418	-0.005	0.018	0.718	0.014	0.010
mean-2	3.528	-0.012	-0.009	0.904	-0.004	-0.014
variance-0	0.036	-0.003	0.041	0.090	0.100	0.095
variance-1	0.222	-0.014	0.004	0.778	-0.002	0.012
variance-2	0.087	0.075	0.110	0.272	0.154	0.026
weight-0	0.034	0.073	0.023	0.051	0.118	0.051
weight-1	0.027	0.020	0.029	0.086	0.143	0.061
weight-2	0.026	0.056	0.086	0.241	0.079	0.052
Interaction Effects						
mean-0	0.181	0.238	0.220	0.375	0.490	0.411
mean-1	11.418	0.084	0.103	0.847	0.087	0.116
mean-2	3.528	0.042	0.025	0.955	0.049	0.033
variance-0	0.036	0.239	0.215	0.384	0.493	0.418
variance-1	0.222	0.066	0.081	0.865	0.064	0.138
variance-2	0.087	0.214	0.272	0.474	0.258	0.186
weight-0	0.034	0.408	0.248	0.309	0.454	0.397
weight-1	0.027	0.253	0.182	0.438	0.604	0.323
weight-2	0.026	0.268	0.178	0.630	0.312	0.171

Table 6.5. Decomposition of variance, GMM. Italian Wikipedia.

variable	variance	rollback_prob r	speed μ	confidence ε	const_succ c_s	const_pop c_p
Truncated GMM, $k = 2$						
Main Effects						
mean-0	1.312	0.075	0.073	0.270	0.133	0.060
mean-1	2.845	0.037	0.032	0.771	0.043	0.014
variance-0	0.533	0.077	0.076	0.281	0.139	0.057
variance-1	0.018	0.019	0.088	0.138	0.097	0.073
weight-0	0.042	-0.006	0.009	0.779	0.105	0.007
Interaction Effects						
mean-0	1.312	0.209	0.384	0.477	0.377	0.088
mean-1	2.845	0.107	0.048	0.866	0.108	0.054
variance-0	0.533	0.214	0.355	0.479	0.375	0.085
variance-1	0.018	0.188	0.423	0.435	0.489	0.182
weight-0	0.042	0.032	0.068	0.805	0.171	0.041
Truncated GMM, $k = 3$						
Main Effects						
mean-0	0.029	0.021	0.019	0.269	0.181	0.101
mean-1	5.127	-0.002	0.010	0.823	0.027	0.018
mean-2	3.444	0.000	0.001	0.848	0.026	0.003
variance-0	0.004	0.034	0.027	0.210	0.125	0.088
variance-1	0.207	0.005	0.012	0.791	0.036	0.003
variance-2	0.033	0.035	0.046	0.400	0.249	0.068
weight-0	0.029	-0.009	-0.015	0.666	0.018	0.080
weight-1	0.009	0.041	-0.003	0.201	0.032	0.085
weight-2	0.020	0.004	0.007	0.571	0.058	0.057
Interaction Effects						
mean-0	0.029	0.252	0.141	0.487	0.382	0.313
mean-1	5.127	0.063	0.062	0.890	0.061	0.078
mean-2	3.444	0.064	0.038	0.925	0.071	0.042
variance-0	0.004	0.284	0.185	0.517	0.403	0.338
variance-1	0.207	0.095	0.067	0.876	0.074	0.056
variance-2	0.033	0.105	0.143	0.515	0.336	0.196
weight-0	0.029	0.152	0.061	0.771	0.119	0.180
weight-1	0.009	0.253	0.164	0.636	0.417	0.279
weight-2	0.020	0.136	0.102	0.721	0.163	0.189

Table 6.6. Decomposition of variance, truncated GMM. Italian Wikipedia.

variable	variance	rollback_prob r	speed μ	confidence ε	const_succ c_s	const_pop c_p
GMM, $k = 2$						
Main Effects						
mean-0	0.511	0.012	0.037	0.294	0.127	-0.014
mean-1	2.323	-0.001	-0.002	0.819	0.007	-0.014
variance-0	0.470	0.018	0.044	0.295	0.144	-0.008
variance-1	0.020	0.001	0.037	0.165	0.043	0.010
weight-0	0.041	-0.009	-0.013	0.793	0.090	-0.002
Interaction Effects						
mean-0	0.511	0.152	0.317	0.534	0.449	0.081
mean-1	2.323	0.065	0.034	0.925	0.102	0.049
variance-0	0.470	0.161	0.280	0.521	0.458	0.070
variance-1	0.020	0.187	0.328	0.570	0.573	0.114
weight-0	0.041	0.030	0.057	0.846	0.166	0.034
GMM, $k = 3$						
Main Effects						
mean-0	0.106	0.051	0.041	0.131	0.075	0.029
mean-1	6.897	0.031	0.009	0.820	0.040	0.005
mean-2	2.856	0.015	0.009	0.910	0.006	0.012
variance-0	0.034	0.061	0.037	0.192	0.019	0.026
variance-1	0.164	0.018	0.011	0.806	0.009	0.018
variance-2	0.058	0.066	0.049	0.331	0.222	0.004
weight-0	0.025	0.017	0.013	0.240	0.043	0.006
weight-1	0.029	0.029	0.041	0.024	0.069	0.020
weight-2	0.020	0.042	0.049	0.370	0.053	0.022
Interaction Effects						
mean-0	0.106	0.163	0.158	0.505	0.580	0.366
mean-1	6.897	0.052	0.056	0.864	0.081	0.079
mean-2	2.856	0.025	0.024	0.954	0.052	0.035
variance-0	0.034	0.155	0.193	0.588	0.524	0.296
variance-1	0.164	0.064	0.073	0.863	0.064	0.135
variance-2	0.058	0.168	0.237	0.468	0.319	0.198
weight-0	0.025	0.229	0.191	0.582	0.545	0.344
weight-1	0.029	0.142	0.113	0.677	0.816	0.193
weight-2	0.020	0.192	0.147	0.746	0.328	0.159

Table 6.7. Decomposition of variance, GMM. French Wikipedia.

variable	variance	rollback_prob r	speed μ	confidence ϵ	const_succ c_s	const_pop c_p
Truncated GMM, $k = 2$						
Main Effects						
mean-0	0.982	-0.005	0.043	0.314	0.140	-0.015
mean-1	2.278	0.015	0.018	0.810	0.027	0.007
variance-0	0.484	0.000	0.050	0.315	0.152	-0.009
variance-1	0.018	0.020	0.073	0.141	0.068	0.040
weight-0	0.040	0.022	0.018	0.779	0.129	0.028
Interaction Effects						
mean-0	0.982	0.145	0.309	0.517	0.479	0.077
mean-1	2.278	0.063	0.037	0.896	0.112	0.050
variance-0	0.484	0.149	0.284	0.509	0.483	0.072
variance-1	0.018	0.188	0.356	0.520	0.622	0.113
weight-0	0.040	0.031	0.066	0.791	0.180	0.038
Truncated GMM, $k = 3$						
Main Effects						
mean-0	0.062	0.056	0.058	0.410	0.065	0.003
mean-1	5.391	0.009	0.002	0.754	0.008	0.002
mean-2	2.826	-0.005	-0.008	0.883	-0.023	-0.013
variance-0	0.020	0.050	0.059	0.317	0.056	0.032
variance-1	0.204	0.007	-0.003	0.735	0.002	0.002
variance-2	0.046	0.078	0.059	0.335	0.206	0.034
weight-0	0.023	0.036	0.014	0.584	0.028	0.013
weight-1	0.013	0.091	0.070	0.083	0.037	0.026
weight-2	0.018	0.026	0.027	0.596	-0.013	0.015
Interaction Effects						
mean-0	0.062	0.188	0.244	0.596	0.266	0.139
mean-1	5.391	0.055	0.120	0.834	0.105	0.097
mean-2	2.826	0.038	0.036	0.964	0.061	0.042
variance-0	0.020	0.165	0.318	0.548	0.346	0.130
variance-1	0.204	0.054	0.121	0.851	0.102	0.114
variance-2	0.046	0.110	0.248	0.448	0.325	0.218
weight-0	0.023	0.176	0.119	0.725	0.193	0.140
weight-1	0.013	0.285	0.267	0.451	0.569	0.244
weight-2	0.018	0.109	0.076	0.829	0.219	0.138

Table 6.8. Decomposition of variance, truncated GMM. French Wikipedia.

variable	variance	rollback_prob r	speed μ	confidence ε	const_succ c_s	const_pop c_p
GMM, $k = 2$						
Main Effects						
mean-0	2.873	0.033	0.068	0.245	0.093	0.043
mean-1	3.885	0.042	0.045	0.736	0.017	0.002
variance-0	0.909	0.042	0.073	0.280	0.096	0.025
variance-1	0.042	-0.003	0.063	0.043	0.001	0.104
weight-0	0.045	0.009	0.018	0.803	0.112	0.020
Interaction Effects						
mean-0	2.873	0.174	0.377	0.515	0.379	0.132
mean-1	3.885	0.116	0.059	0.863	0.124	0.050
variance-0	0.909	0.203	0.336	0.549	0.345	0.118
variance-1	0.042	0.223	0.484	0.405	0.452	0.329
weight-0	0.045	0.033	0.046	0.815	0.147	0.044
GMM, $k = 3$						
Main Effects						
mean-0	1.475	0.008	0.027	0.337	0.022	0.063
mean-1	10.968	-0.012	-0.009	0.858	-0.020	-0.018
mean-2	4.737	-0.014	-0.022	0.861	-0.009	-0.021
variance-0	0.844	0.014	0.039	0.267	0.053	0.067
variance-1	0.238	0.004	0.039	0.459	0.053	0.044
variance-2	0.079	0.036	0.036	0.556	0.019	0.009
weight-0	0.025	0.220	0.061	0.083	0.083	0.073
weight-1	0.026	0.112	0.011	0.254	0.221	0.018
weight-2	0.020	0.042	0.038	0.183	0.067	0.024
Interaction Effects						
mean-0	1.475	0.125	0.335	0.549	0.390	0.149
mean-1	10.968	0.056	0.052	0.923	0.053	0.077
mean-2	4.737	0.063	0.035	0.950	0.062	0.038
variance-0	0.844	0.139	0.367	0.498	0.434	0.139
variance-1	0.238	0.118	0.233	0.646	0.253	0.189
variance-2	0.079	0.152	0.158	0.702	0.171	0.173
weight-0	0.025	0.477	0.241	0.225	0.255	0.443
weight-1	0.026	0.195	0.179	0.544	0.443	0.157
weight-2	0.020	0.279	0.176	0.606	0.408	0.199

Table 6.9. Decomposition of variance, GMM. German Wikipedia.

variable	variance	rollback_prob r	speed μ	confidence ϵ	const_succ c_s	const_pop c_p
Truncated GMM, $k = 2$						
Main Effects						
mean-0	3.114	0.044	0.079	0.291	0.116	0.053
mean-1	3.828	0.033	0.029	0.732	0.039	0.016
variance-0	0.715	0.052	0.075	0.317	0.116	0.036
variance-1	0.040	0.020	0.102	0.019	-0.009	0.108
weight-0	0.043	0.003	0.006	0.793	0.112	0.007
Interaction Effects						
mean-0	3.114	0.181	0.356	0.536	0.358	0.093
mean-1	3.828	0.119	0.062	0.863	0.121	0.050
variance-0	0.715	0.210	0.314	0.569	0.330	0.081
variance-1	0.040	0.238	0.516	0.430	0.481	0.265
weight-0	0.043	0.035	0.051	0.820	0.157	0.042
Truncated GMM, $k = 3$						
Main Effects						
mean-0	1.106	0.059	0.081	0.222	0.094	0.098
mean-1	7.550	0.047	0.012	0.797	0.012	0.009
mean-2	4.506	0.034	0.004	0.835	0.020	0.002
variance-0	0.437	0.059	0.089	0.208	0.104	0.094
variance-1	0.232	0.065	0.108	0.515	0.024	0.018
variance-2	0.061	0.041	0.039	0.500	0.006	0.069
weight-0	0.037	-0.004	0.019	0.761	0.096	0.002
weight-1	0.017	0.020	0.061	0.486	0.016	0.043
weight-2	0.024	0.016	0.078	0.291	0.087	0.019
Interaction Effects						
mean-0	1.106	0.153	0.390	0.458	0.431	0.154
mean-1	7.550	0.100	0.080	0.869	0.059	0.056
mean-2	4.506	0.077	0.041	0.910	0.072	0.038
variance-0	0.437	0.156	0.399	0.441	0.446	0.147
variance-1	0.232	0.137	0.282	0.624	0.209	0.104
variance-2	0.061	0.183	0.171	0.708	0.156	0.224
weight-0	0.037	0.059	0.050	0.810	0.149	0.071
weight-1	0.017	0.148	0.161	0.794	0.244	0.129
weight-2	0.024	0.239	0.183	0.619	0.338	0.178

Table 6.10. Decomposition of variance, truncated GMM. German Wikipedia.

the above procedure both using the variance of each parameter as weight (see column “variance” in tables 6.3–6.10) and without weight, i.e. using $\mathbf{W} = \mathbf{I}$ in eq. (6.2). Surprisingly, the best results are those with no weighting, which are those we choose to report here. In table 6.11 we report the coefficient of determination R^2 for the various choice of the auxiliary model. Figures in bold (only in the rows of the confidence parameter) denote the best auxiliary model for each language. The graphical results of the cross-validation for these models are shown in figg. 6.7–6.10.

Parameter	GMM		Truncated GMM	
	$k = 2$	$k = 3$	$k = 2$	$k = 3$
Portuguese				
Speed μ	0.02	0.03	0.01	0.04
Confidence ε	0.73	0.86	0.70	0.85
Rollback prob. r	0.00	0.16	0.01	0.00
Const. succ. c_s	0.13	0.02	0.36	0.28
Const. pop. c_p	0.02	0.01	0.01	0.01
Italian				
Speed μ	0.00	0.02	0.02	0.00
Confidence ε	0.91	0.90	0.93	0.85
Rollback prob. r	0.01	0.01	0.03	0.00
Const. succ. c_s	0.66	0.30	0.75	0.42
Const. pop. c_p	0.09	0.03	0.01	0.03
French				
Speed μ	0.00	0.01	0.01	0.08
Confidence ε	0.91	0.90	0.76	0.86
Rollback prob. r	0.00	0.03	0.00	0.16
Const. succ. c_s	0.61	0.33	0.69	0.35
Const. pop. c_p	0.04	0.01	0.08	0.09
German				
Speed μ	0.00	0.03	0.01	0.13
Confidence ε	0.91	0.77	0.92	0.67
Rollback prob. r	0.08	0.00	0.06	0.09
Const. succ. c_s	0.50	0.16	0.38	0.27
Const. pop. c_p	0.06	0.01	0.00	0.05

Table 6.11. Results of leave-one-out cross-validation. Coefficient of determination. For each language, the best R^2 attained over parameter ε is shown in bold.

The results of the cross-validation are better than we were expecting. The

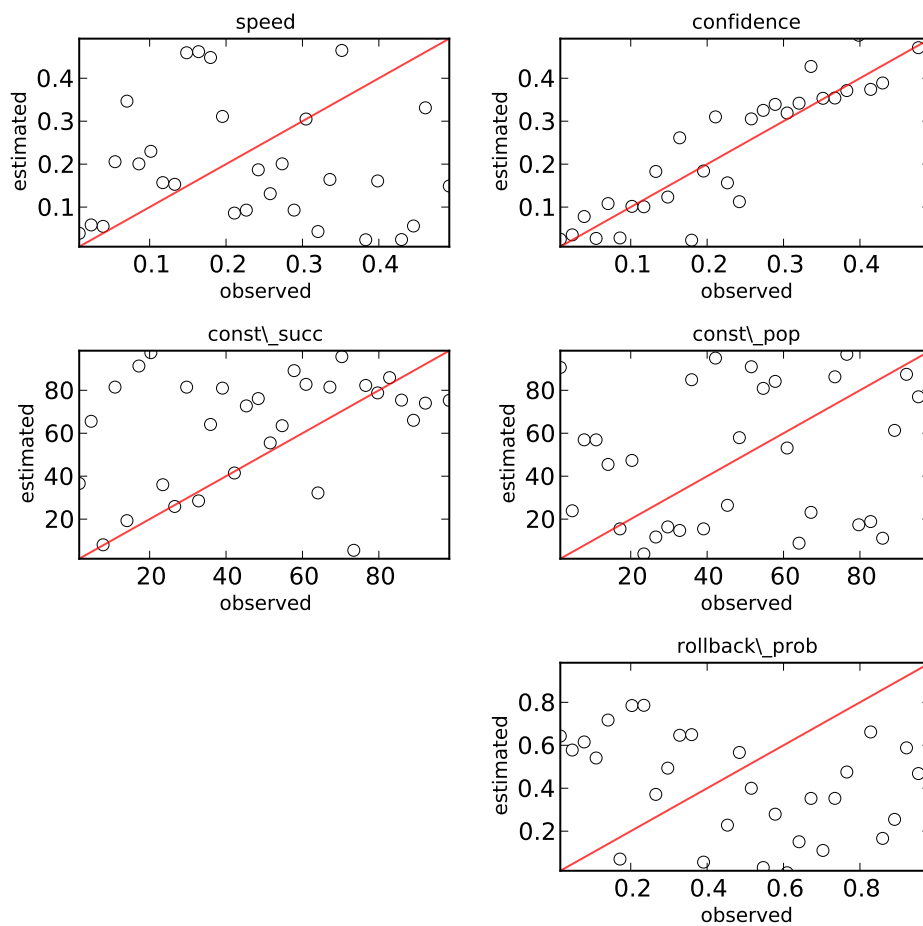


Figure 6.7. Leave-out-out cross validation. Portuguese Wikipedia, GMM, $k = 3$.

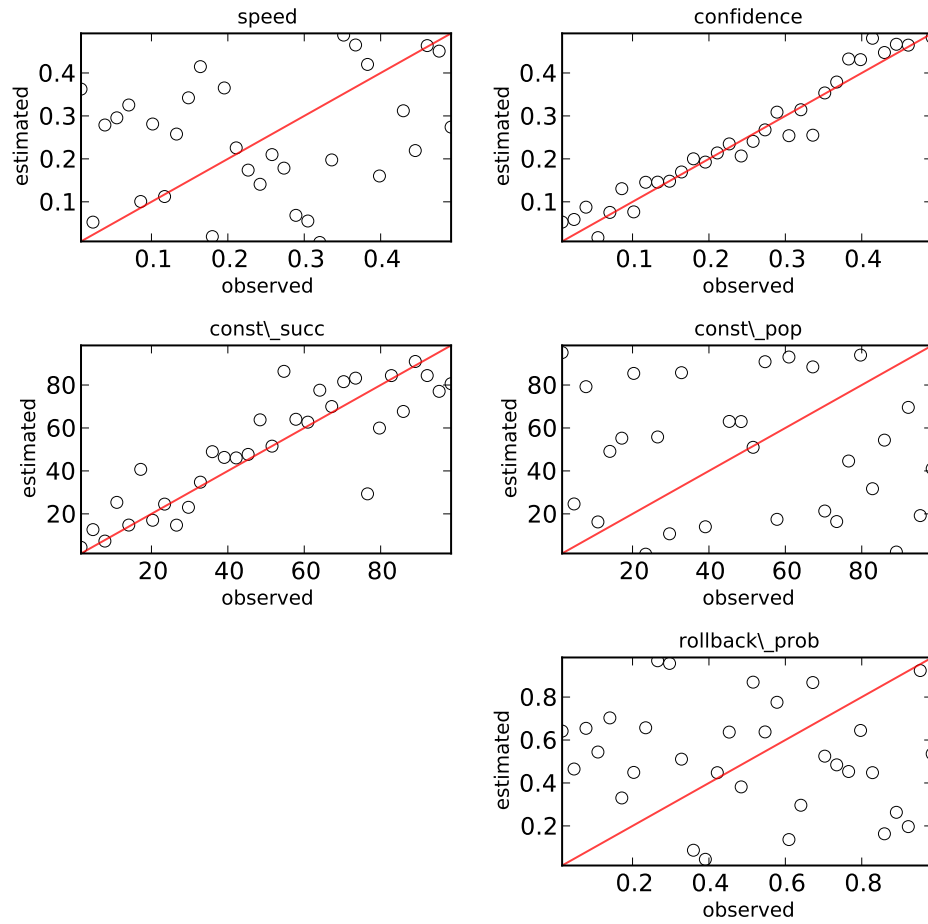


Figure 6.8. Leave-out-out cross validation. Portuguese Wikipedia, Truncated GMM, $k = 2$.

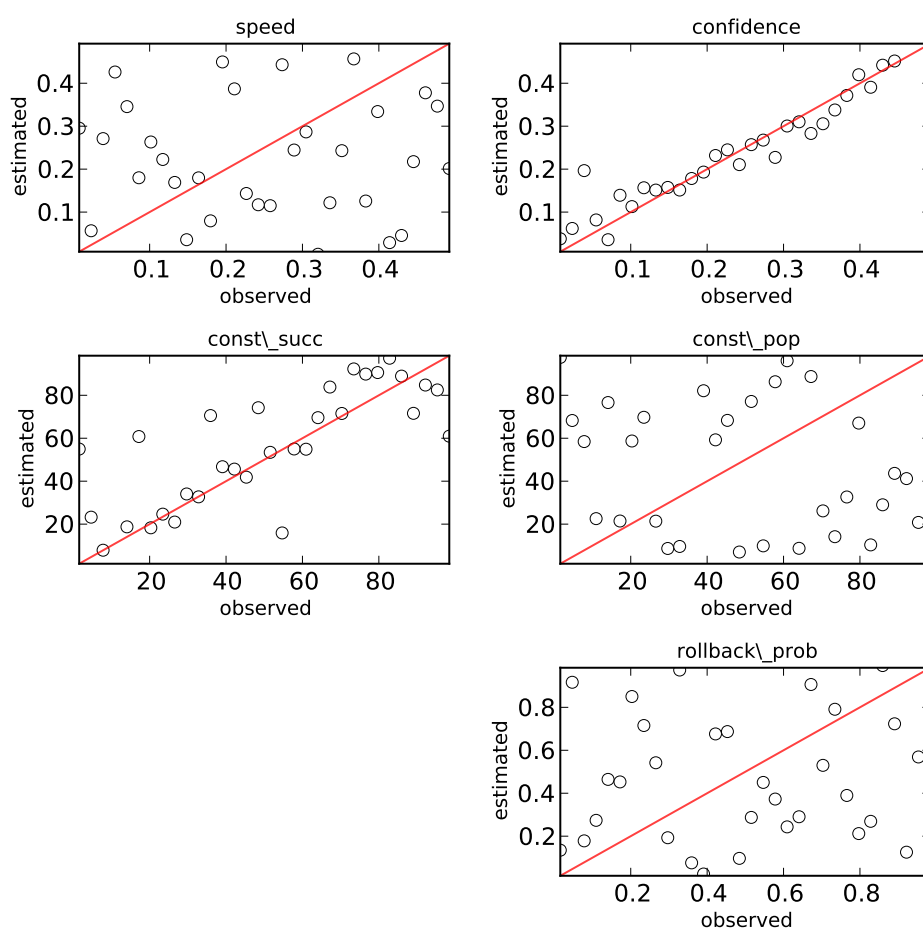


Figure 6.9. Leave-out-out cross validation. Portuguese Wikipedia, GMM, $k = 2$.

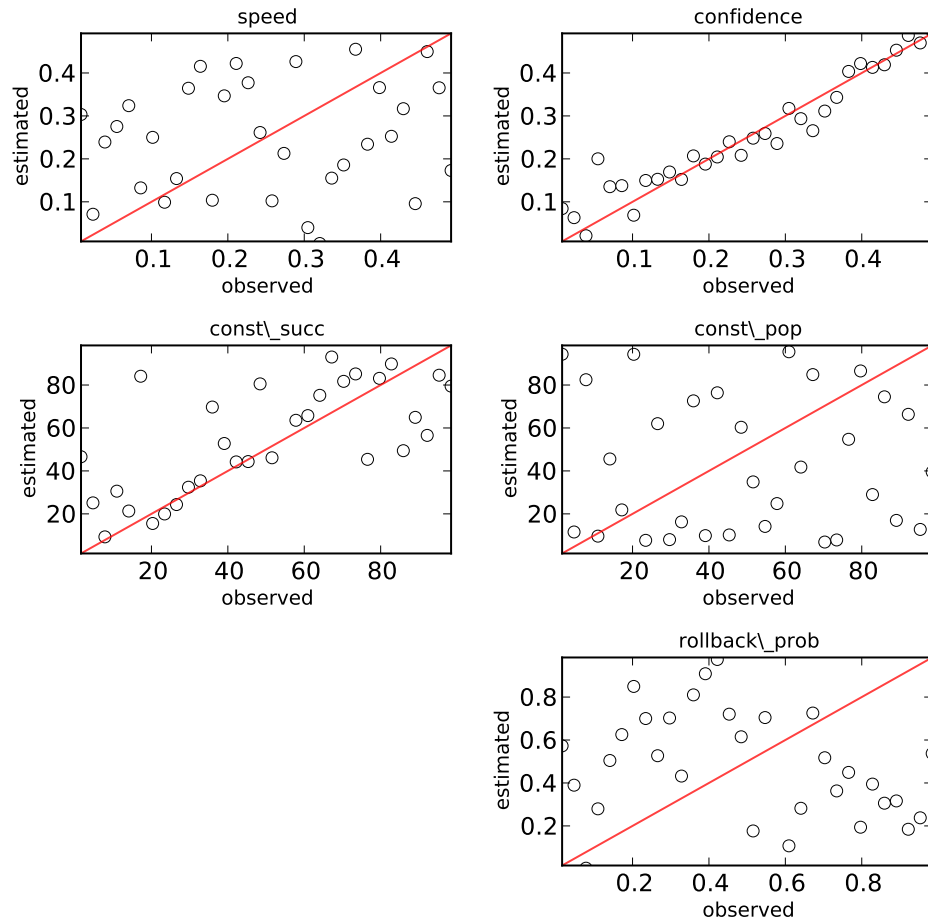


Figure 6.10. Leave-out-out cross validation. Portuguese Wikipedia, Truncated GMM, $k = 2$.

decomposition of variance of the sufficient statistic of the mixture told us basically that the only parameter on which the auxiliary parameters are sensitive is the confidence ε . However in terms of R^2 we see that the indirect inference is moderately accurate in estimating also c_s , the constant term of successes. This can be construed noting that this parameter is able to influence the location of the short-term cluster, a feature that the mixture model is able to detect. The factor screening did not evidence this behavior of c_s because, in the presence of a heavy tailed statistics, changes in this component do not affect much the average activity lifespan of the population. So we see that the indirect inference is indeed an effective estimation technique – provided one chooses the right auxiliary model.

6.6 Results

Having tested the accuracy of the indirect inference technique, we can finally apply it to get estimates of the parameters of our model. We fit our data using the auxiliary models that attain the highest R^2 in estimating ε . See table 6.12 for the results of the calibration.

All other settings are similar to those used for the sensitivity analysis. Standard errors and 95% confidence intervals are computed on a bootstrapped sample with 1000 observations. Figure 6.11 shows the results of the fit, compared to empirical data. We simulate from the calibrated model and plot a kernel density estimate of the synthetic data together with histograms of empirical data.

Parameter	Estimate	Std. Error	95% conf. int.
Portuguese			
rollback_prob	0.52	0.01	0.64
speed	0.46	0.00	0.30
confidence	0.39	0.00	0.08
const_succ	70.78	0.80	49.11
const_pop	51.56	0.79	48.73
Italian			
rollback_prob	0.36	0.00	0.22
speed	0.21	0.00	0.15
confidence	0.49	0.00	0.01
const_succ	53.81	0.61	37.46
const_pop	58.31	0.57	34.93
French			
rollback_prob	0.02	0.01	0.62
speed	0.02	0.00	0.25
confidence	0.49	0.00	0.00
const_succ	3.79	0.86	53.00
const_pop	89.37	0.77	47.69
German			
rollback_prob	0.42	0.01	0.81
speed	0.23	0.01	0.38
confidence	0.49	0.00	0.01
const_succ	1.56	0.79	48.57
const_pop	11.52	1.19	73.21

Table 6.12. Calibrated parameters.

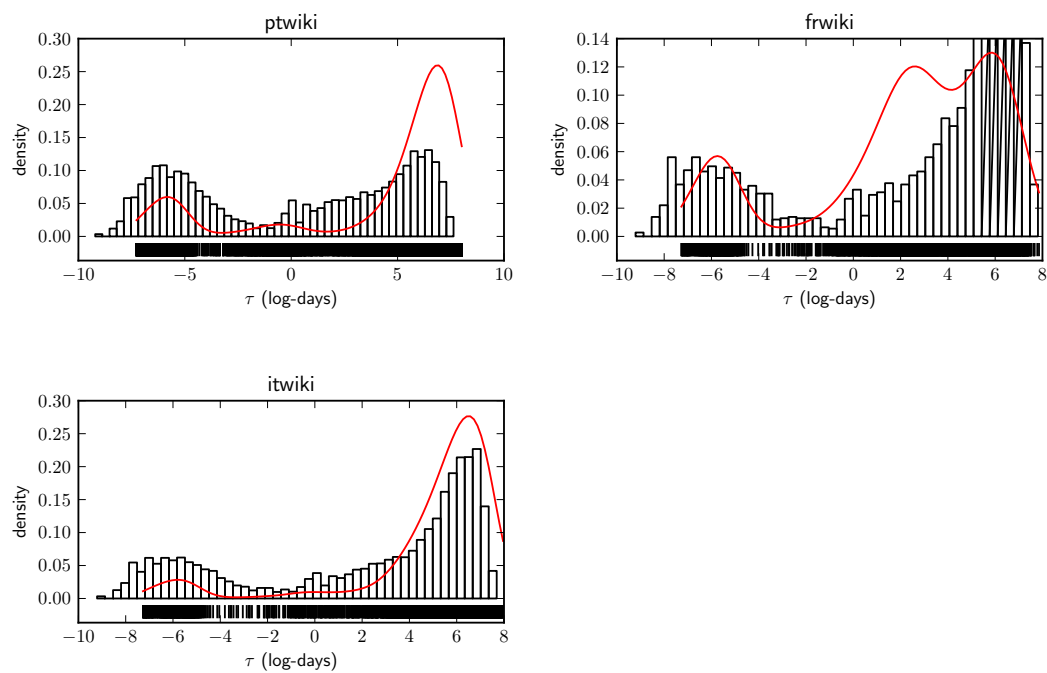


Figure 6.11. Comparison between empirical data (histograms) and simulated data (kernel density estimate, red line), obtained from the calibrated model.

Chapter 7

The life cycle of user activity

7.1 Introduction

How does the editing activity of an editor change through his participation span? In the first version of our peer production model editors were modeled assuming a constant rate of activity, that is, as time homogeneous point processes. Later we revised this assumption, and introduced an editing model with edit cascades. This was needed in order to better describe the bursty nature of editing which, like many human activities happening on the web, is far from being characterized by plain Poissonian statistics¹

But still, in our model it was assumed that users follow the same behavior consistently throughout their whole lifespan. Moreover, it was assumed that all users have the same characteristics in terms of editing activity. Such modeling assumptions seem a fair approximation, at least when taken in the context of modeling how the process of norm formation induces long-term user participation. But outside the context of our model we might ask whether these assumptions are realistic or not: are there instead any distinctive patterns of editing activity that could help explain long-term user participation, user retention, or user departure? And if yes, have these patterns changed noticeably during the history of Wikipedia? In fact, we know that early activity of Wikipedia editors is a good predictor for their long term participation (cf. Panciera et al. [2009]), but what is the general trend in the life cycle of a user after the first month?

Another hypothesis that would be interesting to test is whether editing be-

¹For an explanation involving scale-free statistics, cf. Barabási [2005]; Dezsö et al. [2006]. Also, for an alternative model based on Poissonian patterns, see Malmgren et al. [2008]. Radicchi [2009] found a power-law for the inter-event time distribution of – among others – logging actions by Wikipedia administrators.

havior has changed significantly over the years. Information on the Wikipedia website is organized in different “namespaces”. Namespaces can be thought of as directories: each web-page is identified by a title and by the namespace it belongs to. The purpose of this subdivision is to organize the contents of the wiki website according to the functional and logical criteria. For example, all entries of the actual encyclopedia belong to the so-called “Main” namespace, which in the database is assigned namespace number 0. Personal pages of registered editors belong instead to the “User” namespace (NS 2). Project and community pages are to be found in the “Wikipedia” namespace.

The advantage of this naming convention is that information can be organized according to its function in a natural way. For example, the topic “Vandalism” exists both in the main and “Wikipedia” namespace. In the first case, it is a proper encyclopedic article about the nature of vandalism (i.e. the real-world issue); the latter is instead an internal documentation page about the phenomenon of wiki-vandalism, that is, about any act of vandalism committed on Wikipedia itself.

Related research has so far been interested in characterizing the balance of editing activity among different namespaces only in terms of the historic development of Wikipedia. For example Beschastnikh et al. [2008] analyzed the growth of policy adoption in editorial decisions. Thus, different editing behaviors might be reflected in a different balance of activity among different namespaces.

The research presented in this chapter was performed at the Wikimedia Foundation (<http://www.wikimediafoundation.org>) as part of the 2011 Summer of Research initiative.

7.2 Methods

We use editing activity on pages as a general measure of user participation. We measure it as number of contributions (i.e. edits) per day. The rate of activity is a less problematic measure than the lifespan of user activity, i.e. the time elapsed between the first and the last edit of a user, for two reasons. First, users may take long breaks between active periods, and this makes the overall lifespan difficult to interpret. This problem does not have a direct solution, as the sequence of revision metadata lacks any indicator that tells us if a user is still involved in the project (actively) during any such long break.

Second, activity lifespan might be inaccurate, since retired editors sometimes perform edits even long after they stopped contributing. These transitory come-

backs might happen for reasons other than pure contribution, for example to respond to somebody on their talk pages. This means that the activity span may not faithfully reflect the actual period of activity of a user, which is instead better described by the editing rate.²

7.2.1 Data Collection

We extracted data from the Toolserver replica of the English Wikipedia database in August 2011.³ Our data is actually the metadata attached to each revision performed on any Wikipedia page, and comprises user name, timestamp, page title, and page namespace. We selected only registered users which, at the time we collected the data, amounted to $N = 3,484,55$. We grouped revision timestamps by user and then ordered the timestamps of each user chronologically. Let us consider the i -th user, who did n_i edits, at times $t_1^{(i)}, t_2^{(i)}, \dots, t_{n_i}^{(i)}$. His overall (i.e., average) editing activity rate is defined as the total edit count n_i divided by the total activity lifespan:

$$a_i = \frac{n_i}{t_{n_i}^{(i)} - t_1^{(i)}}. \quad (7.1)$$

Our idea is to perform a longitudinal study of the life cycle of user activity. This means that, instead of treating each user individually, we group them in homogeneous groups, and study the average behavior of users within a specific age group and total activity level. This is motivated by two things. First, a longitudinal methodology is suited for tracking the changes in editor activity over the history of Wikipedia. The second motivation is that estimating the daily rate of edits of an individual user can be difficult due to the presence of noise both at the individual and population level.

Regarding this last point, in fact, we note two things. At the individual level, noise is due to the bursty nature of the editing activity. Most human activities shows strong circadian and weekly rhythms. In general, online activity is not immune to this; email usage is an example of this (cf. Malmgren et al. [2008, 2009]). In order to reduce the level of noise at the individual level, we compute the daily average rate of contributions of a user over a 30 day period, or over

²It is customary for editors of the English Wikipedia to announce their retirement from the scenes on their user page using a special code, which the wiki engine displays as a black sign. See <https://en.wikipedia.org/wiki/Template:Retired>. For the current list of editors who declared their retirement in this way, see instead <https://en.wikipedia.org/wiki/Special:WhatLinksHere/Template:Retired>.

³<http://toolserver.org>.

the total user lifespan, if it is less than 30 days. Second, at the population level we find that activity levels across users are highly heterogeneous, as some users are naturally more productive than others. We will actually quantify this later.

Cohort definition

Our objective is to perform a longitudinal study of user activity over the community of Wikipedia, which means that we need to group users into different cohorts. Cohorts are formed in order to group together users who:

1. have similar productivity levels, thus giving a better estimate of the daily rate of editing activity,
2. begin their participation period more or less at the same time, thus controlling for other endogenous and exogenous factors such as level of tolerance in the community, etc.

We define cohorts as groups of users who become active in the same period and whose editing activity rate (7.1) is within a certain range. The choice of this period is arbitrary, and different periods will give results that differ quantitatively from each other. In practice, a cohort should have enough users to yield decent statistical estimates, but should not lump together users who are subject to different social and environmental factors. As a trade-off between these two criteria we set the period associated to each cohort to be equal to 30 days.

To track the beginning of the active period of a user we use the timestamp of the first revision ever performed. We use the date of first edit instead of account registration because the distribution of the lag between registration and first edit of users happens to be markedly skewed (see below), and thus the timestamp of registration might inaccurately reflect the beginning of activity of a user.

Of course we should note that the global rate is defined in terms of the total lifespan $t_{N_i}^{(i)} - t_1^{(i)}$, and thus our cohort definition is in principle likely to show the same inaccuracy issue discussed above. However, we can assume that these inaccuracies affect users uniformly at random, so using 7.1 should not introduce significant biases in the definition of user cohorts.

We bin activity levels of users logarithmically, in base 10. As a reference value, a global activity of 10^{-4} edits/s (or simply s^{-1} , which we use interchangeably throughout the rest of the chapter) corresponds to $n_{\text{month}} = 259.2$ edits in a month. So, our classes roughly corresponds to 2.5–25 edits per month, 25–250 edits per month, 250–2,500 edits per month, etc.

Since we want to track the activity rate as a function of time since the inception of the active period, all user histories are, within the same cohort, shifted so that the date of first edit of the account corresponds to the origin (i.e. $t = 0$) of the x -axis.

Lag between registration and first edit

Even though we had access to the date of registration, we decided to define our cohorts using the day in which users perform their first edit. We used the day of first edit instead of registration because it is a more accurate measure of the beginning of the active period of an editor. Figure 7.1 shows that a significant portion of users (about 20%) register but do not perform their first edit within the same day. Instead, the distribution of lags between registration and first edit appears to show a heavy tail, with several instances past 1 month and even past 1 year. Together with this, we also see a markedly bimodal behavior. The plot depicts the mixture components as stacked area plots. This means that for any value of the registration lag (x -axis), the value of the density of each component of the mixture is given by the height of the respective colored area at that point.

Data quality

Finally, since we want to study human contributors only, we also filter out robot accounts. In particular, we take three measures to remove them: first, we remove all accounts in the list of “flagged” bots, that is, the official community list of robot accounts.⁴ Second, we remove all unflagged accounts that have been identified as being run by a script. Third, we further restrict our sample only to users with at least 2 edits and with a lifespan longer than 1 hour.

7.3 Results and discussions

7.3.1 Editor productivity

Before analyzing any possible variation of the editing activity rate over time, we explore the variability within our sample. Wikipedia editors are known to show a high degree of heterogeneity in terms of contributions, as the edit count distribution (i.e. the total number of contributions performed by an editor)

⁴The term comes from the presence of a boolean field in the user database – a “flag” in the technical jargon – that specifically signals whether an account is operated by a human or by a computer program.

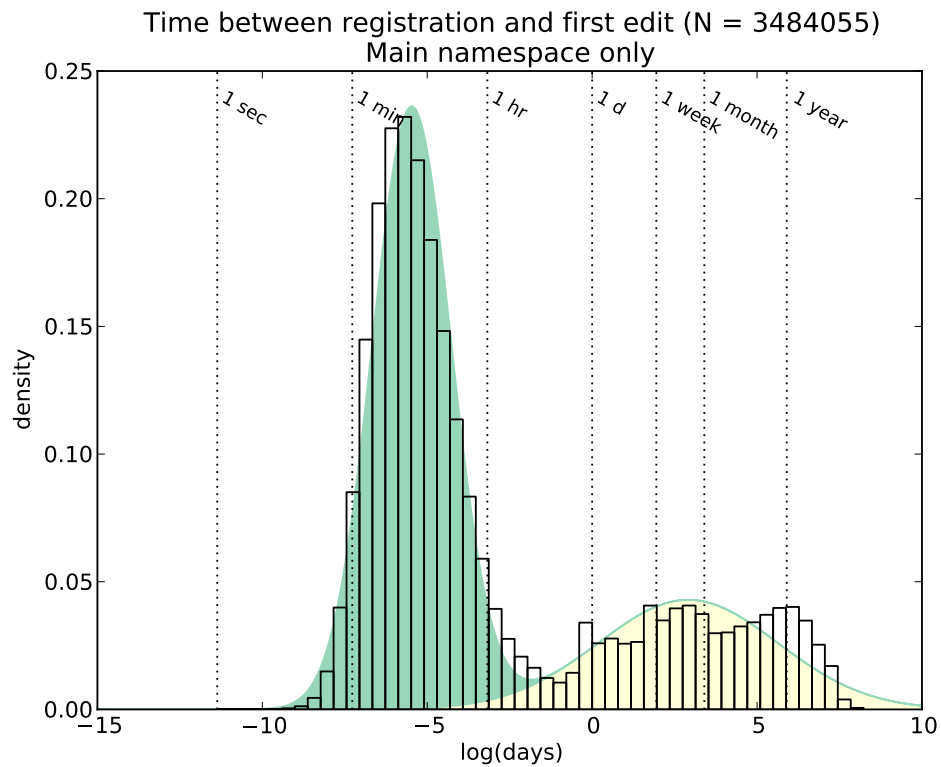


Figure 7.1. Distribution of lags between registration and first edit, for editors who performed their first edit in the main Namespace (encyclopedic articles) (histogram: empirical data). A Gaussian mixture model is fit to the empirical data (stacked area plots).

presents a marked heavy tail (Ortega and Gonzales-Barahona [2007]). However, the sample contains user accounts of different ages, so this heterogeneity might just be an effect of mixing the different lifespans of users, and not due, in the main, to different levels of productivity.

We therefore analyze how editing activity a changes as the total edit count n of a user grows. Figure 7.2 depicts the relationship between the two variables; it reproduces a similar plot found in the work of Radicchi [2009] on human activity on the Web. Radicchi obtained his plot using a related, but smaller, dataset: a snapshot of the “logging” table from 2008.⁵ Figure 7.2 is instead obtained from the contents of the “revision” table as of August 2011, and thus contains activity from the whole population of users, and not just administrator users.

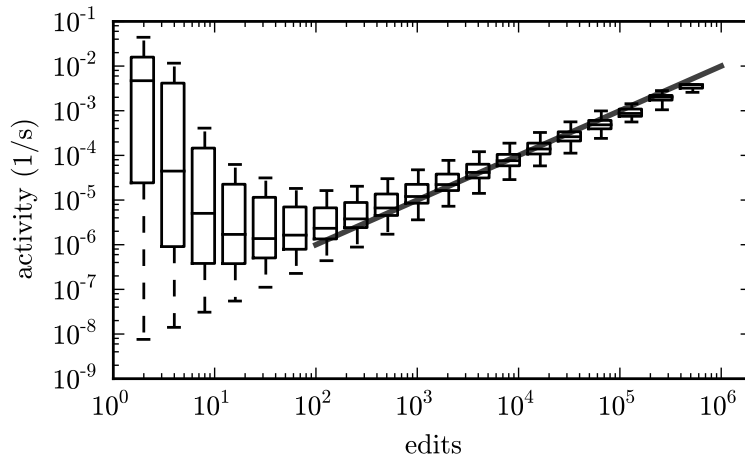


Figure 7.2. User activity (in s^{-1}) versus editing level (number of edits). Users are logarithmically binned (base 2) according to the number of edits performed. For each of these samples, a box-plot shows the distribution of editing activity. Whiskers: 10th–90th percentiles; boxes: 25th–75th percentiles; line: median. The black line (slope = 1) is a reference for the reader to show the linear relationship.

In the plot, boxes stretch between the 25th and 75th percentile, whiskers correspond to the 10th and 90th percentile, and the line inside boxes represents the 50th percentile (i.e. the median). Users are binned logarithmically (in base 2).

⁵This table records all bookkeeping actions performed by administrators, such as page re-naming, archiving, or deletions, etc.

The plot shows that, for $n \gtrsim 200$, a linear relationship (in log-scale) holds between number of edits and activity. However, if we take a given bin of activity greater than $a \geq 10^{-5}$ edits/s, that is, if we set for a given interval on the y-axis greater than 10^{-5} edits/s, and scan the graph horizontally to see where we make any observation, we see that there are two categories of users, those with low edit counts (on the left), who therefore must have performed such edits in a relatively short time frame, and those with high edit counts (on the right), whose span of activity is considerably longer. This is consistent with the fact that the distribution of lifespan is bimodal, as we saw in chapter 3, but also implies that users are indeed marked by different rates of productivity since the very early stages of their careers.

7.3.2 User activity trends

We now look at how editing activity changes over time. The first, qualitative, result is that editing activity evolves over time differently depending on user productivity. For low activity users ($a < 10^{-5}$ edits/s, roughly less than one edit per day), the peak of activity happens in the first 30 days. For high activity users (more than one edit per day), productivity peaks much later. For example, Figure 7.3 shows four cohorts, all from January 2006:

The dashed lines refer to a non-parametric fit performed with a cubic splines model. The smoothness factor of the splines was determined using cross-validation. Vertical lines mark the day of peak activity predicted by the spline model. For editors in the top plot, that is, cohorts with $10^{-4} \text{ s}^{-1} \leq a < 10^{-3} \text{ s}^{-1}$ (in the legend labeled with “-4”) and $10^{-5} \text{ s}^{-1} \leq a < 10^{-6} \text{ s}^{-1}$ (legend label: “-5”) this day occurs after (roughly) 450 and 300 days. For the bottom plot, the peak of activity occurs in the first 30-days period – presumably on the very first day.

For low-activity cohorts a surge in activity is also noted towards the end of the observation window. However, error bars for those measurements (not shown in the plots) also grow dramatically. This is a consequence of the low number of editors still active at such a late stage, which makes the estimates of the average activity rate very noisy. In the rest of the analysis we decided to filter out observations based on a criterion of dispersion: we excluded those estimates of the average rate whose signal-to-noise ratio was below a threshold. We found that a value of 10 gives good results without excluding too many observations.

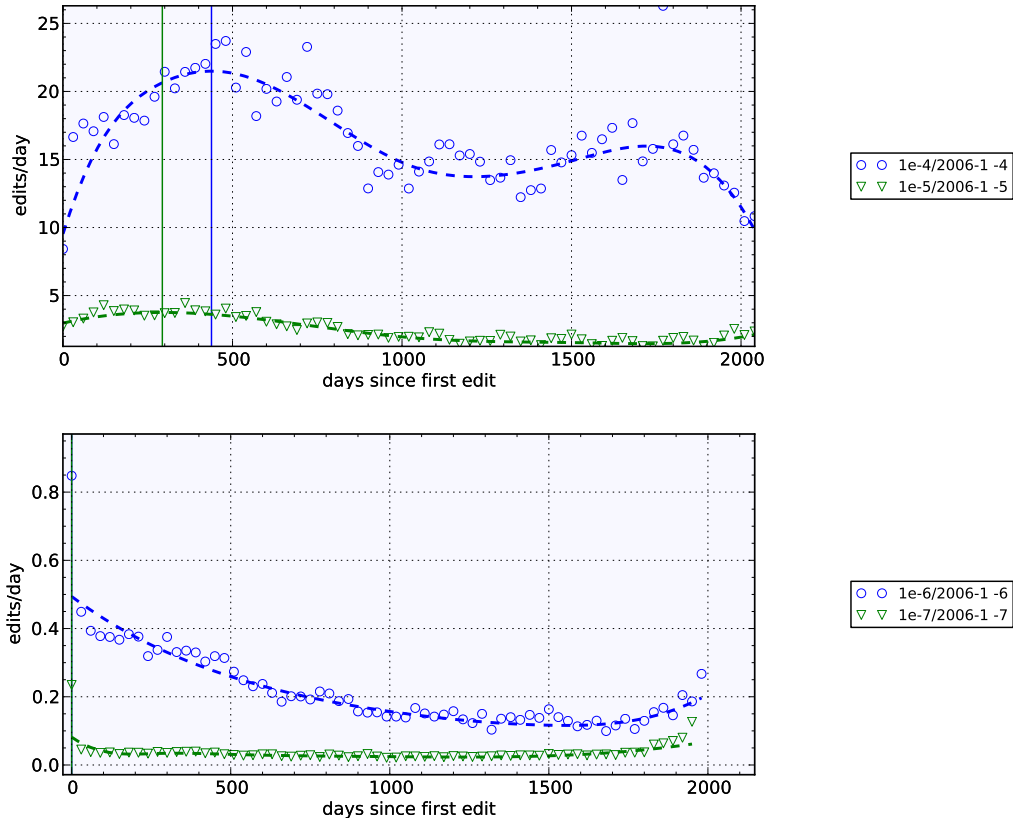


Figure 7.3. Daily edit rate a as a function of time since first edit, for Wikipedia editors who did their first edit in January 2006. Top: $10^{-4} \text{ s}^{-1} \leq a < 10^{-3} \text{ s}^{-1}$ (blue dots), $10^{-5} \text{ s}^{-1} \leq a < 10^{-4} \text{ s}^{-1}$ (green triangles). Bottom: $10^{-6} \text{ s}^{-1} \leq a < 10^{-5} \text{ s}^{-1}$ (blue dots), $10^{-7} \text{ s}^{-1} \leq a < 10^{-6} \text{ s}^{-1}$ (green triangles).

7.3.3 Time scale of activity decay

We also tried parametric models and preliminary results suggest that the post-peak decay in editing activity by users with $a < 10^{-5} \text{ s}^{-1}$ follows a stretched exponential law. Figure 7.4 shows, as an example, the fit of the average activity rate of three cohorts with activity level $10^{-6} \text{ s}^{-1} \leq a < 10^{-5} \text{ s}^{-1}$ from the month of January of three consecutive years: 2006, 2007, and 2008. We fit these data using least squares, that is, we minimize:

$$\chi^2 = \sum_{i=1}^k \left(\frac{y_i - y(t_i; a, \tau, \beta, c)}{\sigma_i} \right)^2 \quad (7.2)$$

where y is a stretched exponential function:

$$y(t; a, \tau, \beta, c) = a \exp \left(- \left(\frac{t}{\tau} \right)^\beta \right) + c \quad (7.3)$$

Under the hypothesis that data have been generated by the model in 7.3 by the effect of Gaussian noise, 7.2 should follow the χ^2 distribution with $\nu = k - 4$ degrees of freedom. Thus we compute, as goodness-fit-measures, the p -value associated to the fit; we also test for normality of the residuals using the K^2 omnibus statistics (D'Agostino et al. [1990]). Finally, we compute the coefficient of determination. Results for the fit for the three cohorts are reported in table 7.1. Since $\chi^2 \approx \nu$ for 2007 and 2008, we can conclude that in these two cases we have moderately good fit to the data. For 2006, instead, we see from the p -value a clear rejection of the stretched exponential hypothesis. This pattern seems to hold also in general, i.e. for the whole dataset.

So, only recent cohorts seem to reliably support the stretched exponential model. The reason for this could be twofold. First, very early cohorts simply don't have enough observation: in fact, for the years before 2005, there are simply too few editors to provide reliable estimates of the average rate of editing. In the period between the end of 2005 to early 2007, Wikipedia underwent a dramatic growth. The sudden growth might not interact well with our cohorts period of 1 month, or it could be that too many users registered and performed very few edits, just out of curiosity. This might explain the fact that the fit for the Jan. 2006 cohort underestimates the initial drop of activity, an effect that can be further appreciated comparing the residuals (fig. 7.4, bottom panel) in the blue curve with the other two.

Cohort	χ^2	d.f.	p -value	K^2	p -value (resid.)	R^2
Jan. 2006	197.39	47	0.00	8.85	> 0.01	0.94
Jan. 2007	43.93	34	0.12	2.82	0.24	0.97
Jan. 2008	17.54	19	0.55	3.69	0.15	0.97

Table 7.1. Stretched exponential fit. Goodness-of-fit test. Significant values of the χ^2 are shown in bold. Cohorts

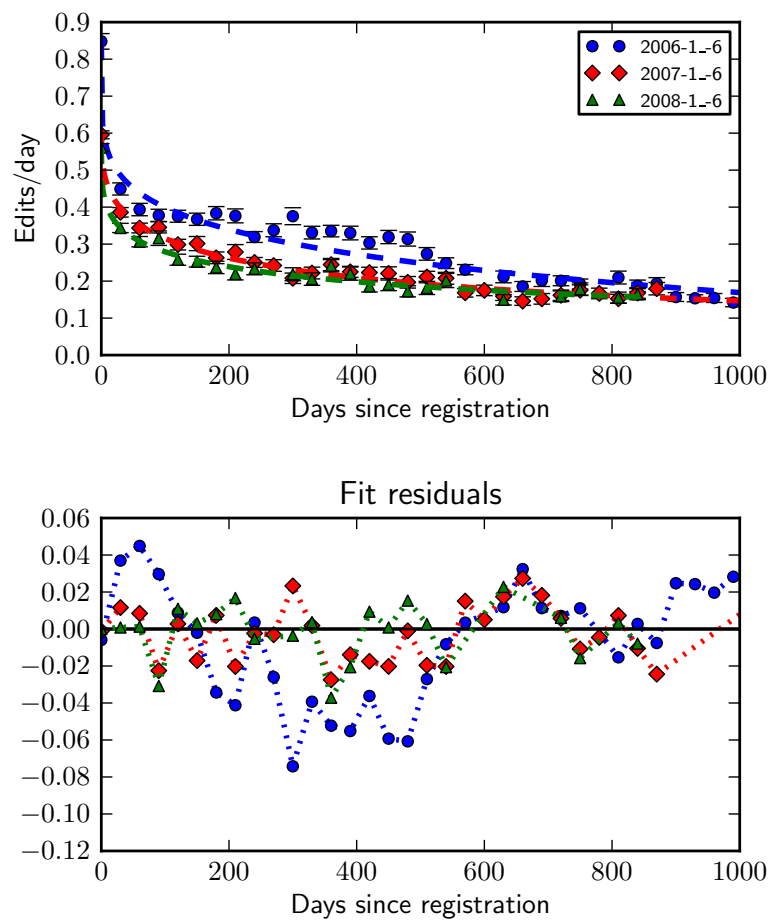


Figure 7.4. Stretched exponential fit. Average activity rate in three cohorts from January 2005 (blue circles), 2007 (red diamonds), and 2009 (green triangles). Activity level $10^{-7} \text{ s}^{-1} \leq a < 10^{-6} \text{ s}^{-1}$. Top: data and fit. Bottom: residuals. Error bars represent the standard error of the mean.

7.3.4 Community dynamics

To get an idea of how editor activity has changed since the inception of the English Wikipedia we can compute the location and the value of the peak of activity of each cohort, and plot it as a function of the cohort month. Such plot gives us the evolution of the productivity of users. By plotting different activity levels, we see the evolution of editing activity going on in the community broken down by productivity levels. These plots are shown in figures 7.5 and 7.6.

For high activity users, i.e. those with $a > 10^{-5} \text{ s}^{-1}$, (more than one edit per day), they show that recent cohorts are reaching their peak of activity earlier than the early ones. Moreover at peak activity they are less productive than those users that, having the same level of productivity, joined in previous years.

For low activity users, the plot is not clear enough because almost invariably these groups have their peak of activity within the first 30 days, so the peak date in the graphs is always equal to zero. This doesn't give a clear picture of the dynamics at this range of activity. To better understand the dynamics at low activity ranges, we can exploit the fact that, at these ranges, activity follows a stretched exponential decay, and compute the mean relaxation time $\langle \tau \rangle$, which is essentially given by the integral under the activity curve.

In fact, for the model of 7.3 $\langle \tau \rangle$ can be computed as:

$$\langle \tau \rangle = \int_0^\infty dt \exp\left(-\left(\frac{t}{\tau}\right)^\beta\right) = \frac{\tau}{\beta} \Gamma\left(\frac{1}{\beta}\right) \quad (7.4)$$

where Γ is the gamma function. $\langle \tau \rangle$ is a measure of how much time it takes for the whole group to reach their baseline activity.

The dynamics of the mean relaxation time is shown in fig. 7.7. These plots show that even low productivity users are losing participation momentum faster than their counterparts from the earlier stages of the project. The only notable exception to this is the group $10^{-8} \text{ s}^{-1} \leq a < 10^{-7} \text{ s}^{-1}$, which is fairly stable, meaning that participation at this (very low) level of activity is still fairly stable. This is probably related to the fact that editors in this group edit only occasionally and thus are not really affected by changes in the structure of the community.

These plots suggest that there is an aging effect going on in Wikipedia, that is, the cycle of user participation is getting shorter as time passes, since editing activity for newer cohorts peaks earlier than in older cohorts. Another way of visualizing this effect comparing different activity ranges, is to rescale each curve displayed in figures 7.5 and 7.6 by its average peak activity. In practice we take a given level of productivity, compute the peak activity for all the cohorts starting

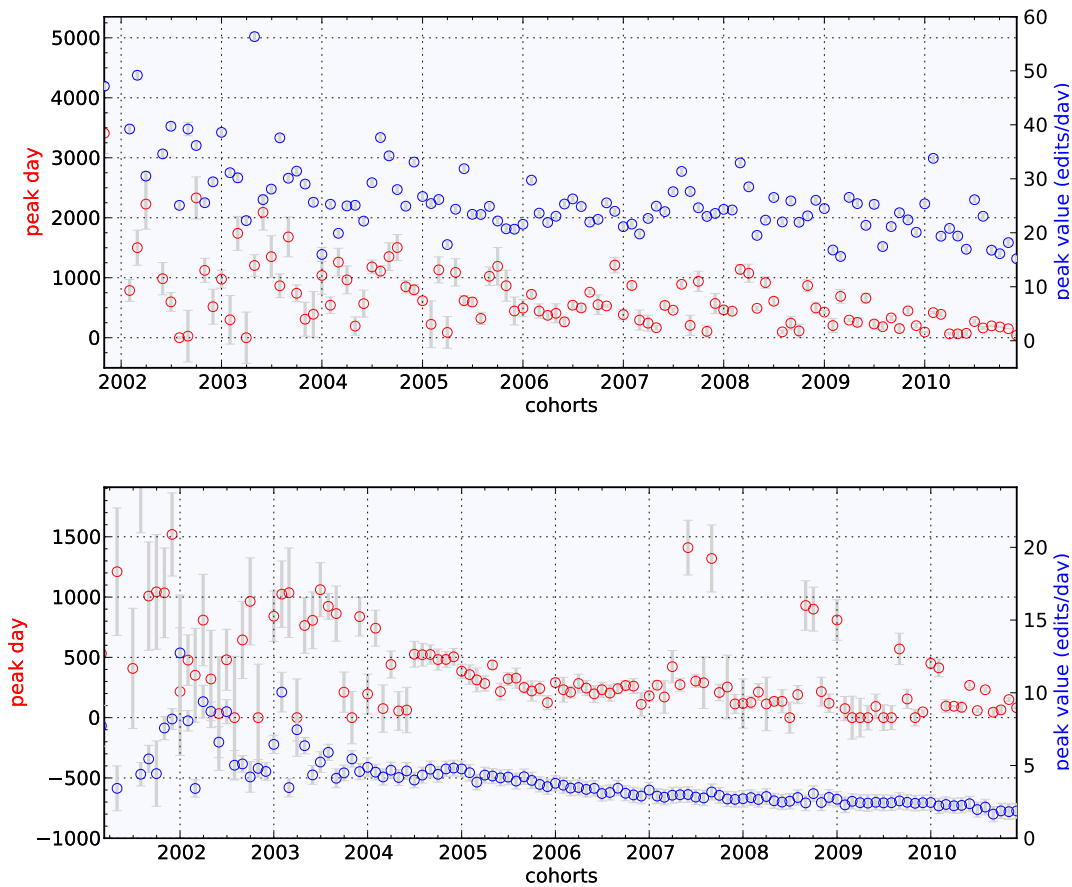


Figure 7.5. Evolution of editing activity over the history of Wikipedia. Peak date (red circles, right y-axis) and peak value (blue circles, left y-axis) of edit activity rates of cohorts for different activity levels. Top: $10^{-4} \text{ s}^{-1} \leq a < 10^{-3} \text{ s}^{-1}$. Bottom: $10^{-5} \text{ s}^{-1} \leq a < 10^{-4} \text{ s}^{-1}$.

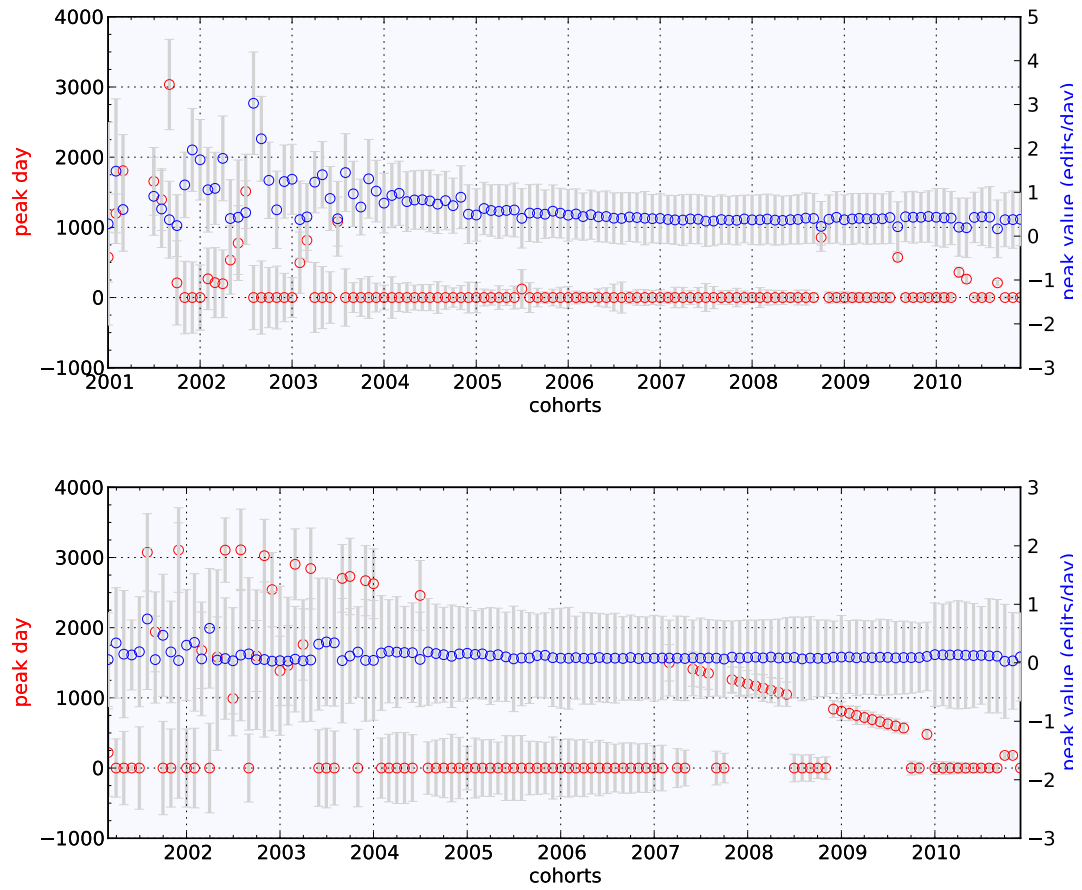


Figure 7.6. Evolution of editing activity over the history of Wikipedia. Peak date (red circles, right y-axis) and peak value (blue circles, left y-axis) of edit activity rates of cohorts for different activity levels. Top: $10^{-6} \text{ s}^{-1} \leq a < 10^{-5} \text{ s}^{-1}$. Bottom: $10^{-7} \text{ s}^{-1} \leq a < 10^{-6} \text{ s}^{-1}$.

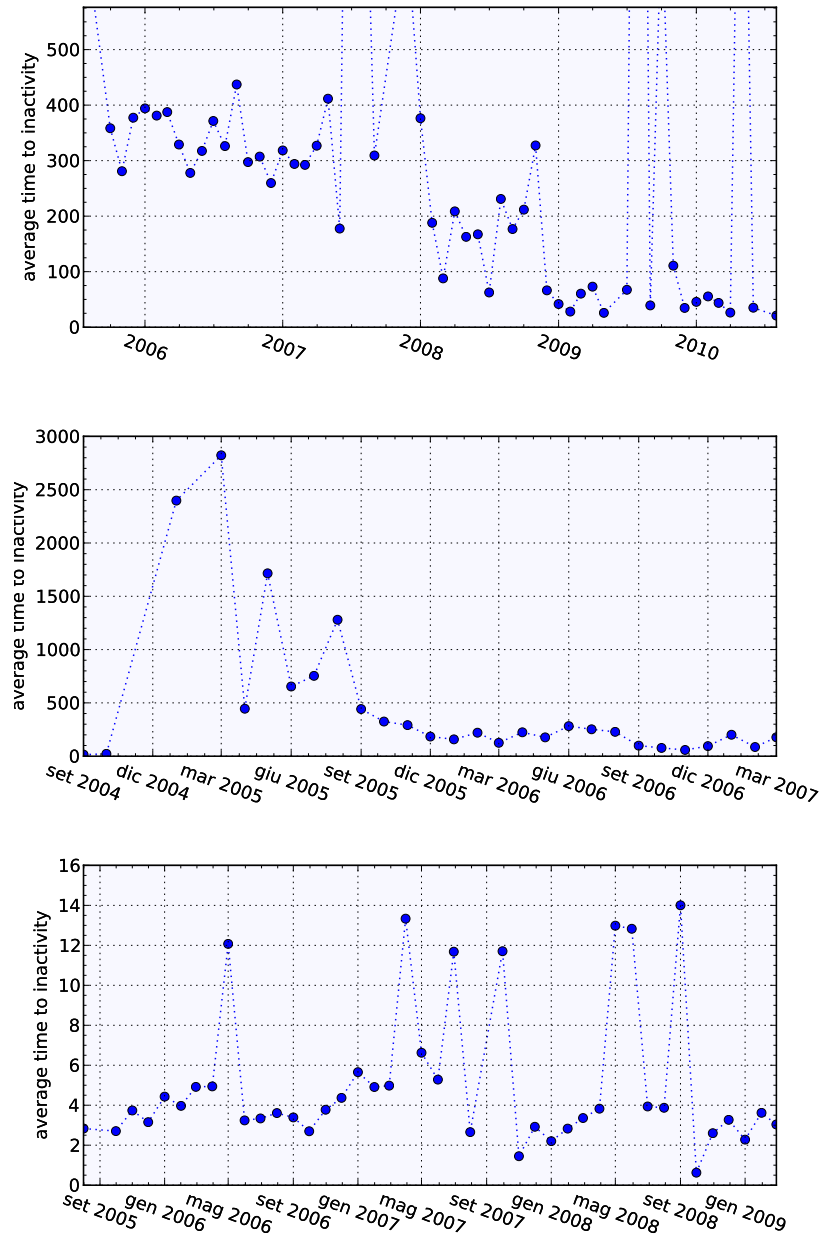


Figure 7.7. Evolution of editing activity over the history of Wikipedia. Mean relaxation time (blue circles, in days) of editing activity of cohorts for different activity levels. Top: $10^{-6} \text{ s}^{-1} \leq a < 10^{-5} \text{ s}^{-1}$. Middle: $10^{-7} \text{ s}^{-1} \leq a < 10^{-6} \text{ s}^{-1}$. Bottom: $10^{-8} \text{ s}^{-1} \leq a < 10^{-7} \text{ s}^{-1}$.

from the inception of Wikipedia, and divide these by the average value. This has the effect of rescaling all values, allowing the dynamics of groups with different productivity levels to be compared. In other words, we can see in this way how fast the slowdown is occurring at different levels of activity. Figures 7.8, 7.9, 7.10, and 7.11 show the results of this comparison. To avoid clutter, we plot together the high activity cohorts ($a \geq 10^{-4}$ edits/s), and those with low activity ($a < 10^{-4}$ edits/s).

In these plots data are further broken down by namespace, that is, we disaggregate edit counts by namespace of destination and then proceed as usual with computing the average rates, fitting our non-parametric spline model, and finding the peaks. We report this break down for the following pairs of namespaces:

- All namespaces.
- “Main” and “Talk” (NS: 0, 1).
- “User” and “User_Talk” (NS: 2, 3).
- “Wikipedia” and “Wikipedia_Talk” (NS: 4, 5).

The first thing to note is that the slowdown is not occurring uniformly over the whole range of the editing activity: if we look at the peak dynamics for high activity users in all namespaces we see that the group $10^{-6} \text{ s}^{-1} \leq a < 10^{-5} \text{ s}^{-1}$ are slowing down steadily, compared to the other two groups. This range of activity corresponds roughly to people performing between 25 to 250 edits per month, whereas people in the two other groups do $10\times$ and $100\times$ that amount of edits. Similar trends hold when looking only at edits to namespaces Main and Talk. This is to be expected, given that most user activity is concentrated into these two namespaces.

For low activity classes ($a < 10^{-6}$ edits/s) we see yet different trends. The group doing between 2 and 25 edits/month seems to slow down whereas the groups respectively 10 and 100 times less active seem not to experience any notable slowdown – which, again, is expectable, given that these users are just occasional users.

In conclusion, our technique allows us to understand how editing activity is changing over time in the whole community, and especially how this change is occurring for different groups of users: very active users seem to have kept up with their pace essentially unchanged since the inception of Wikipedia; other groups, instead, seem to undergo shorter and shorter turnovers. These are important findings to understand how activity is changing over a stratified and heterogeneous community such as the English Wikipedia.

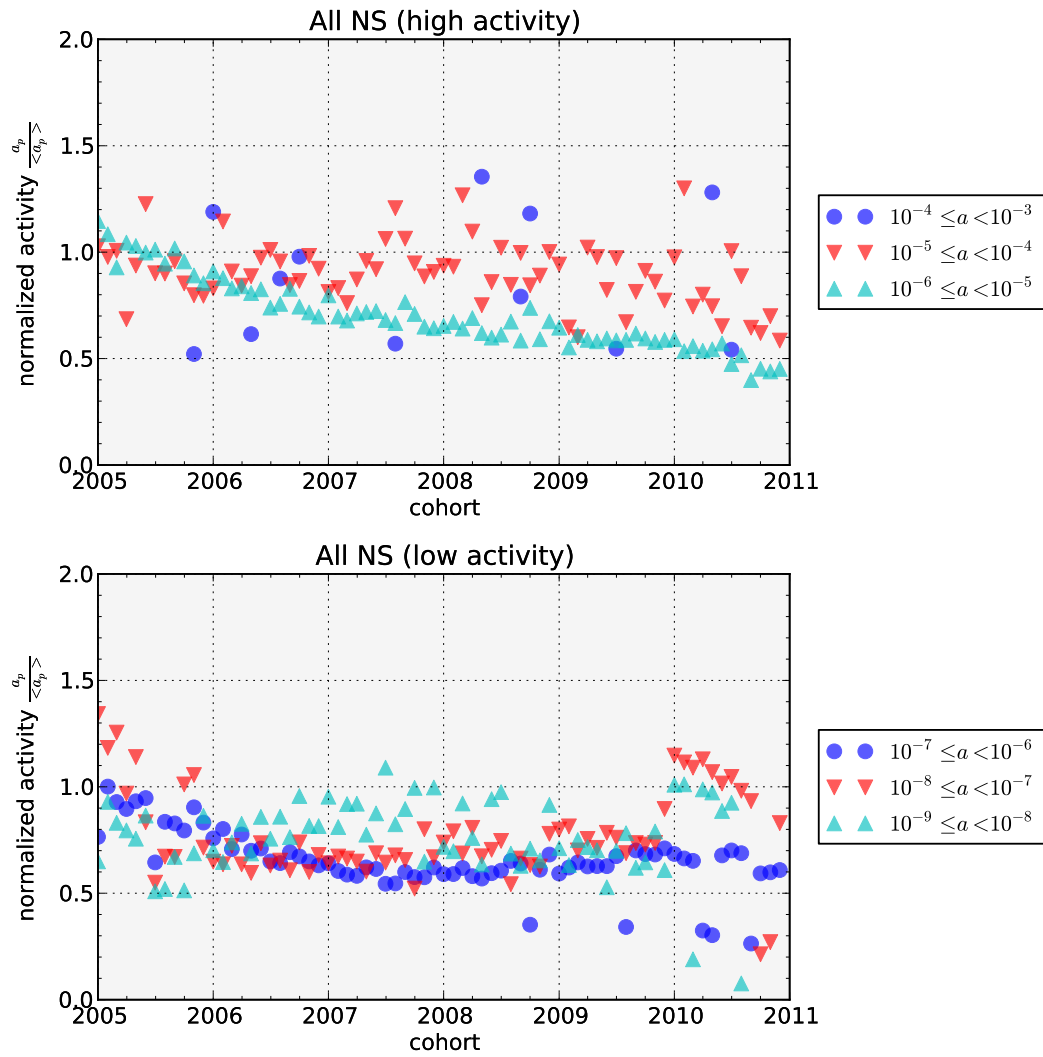


Figure 7.8. Normalized activity rate over time, edits to all namespaces. Top: high activity cohorts. Bottom: low activity cohorts.

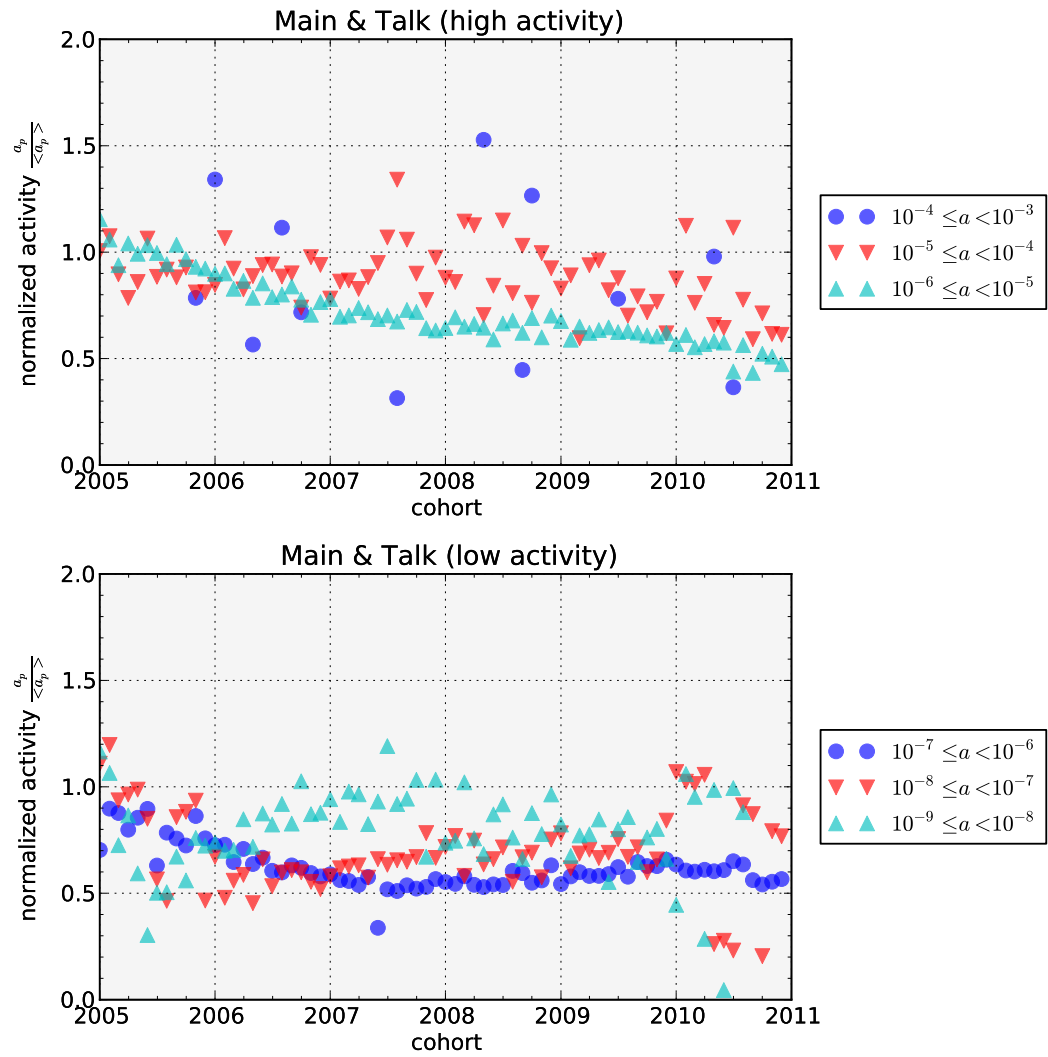


Figure 7.9. Normalized activity rate over time, edits to NS 0 & 1. Top: high activity cohorts. Bottom: low activity cohorts.

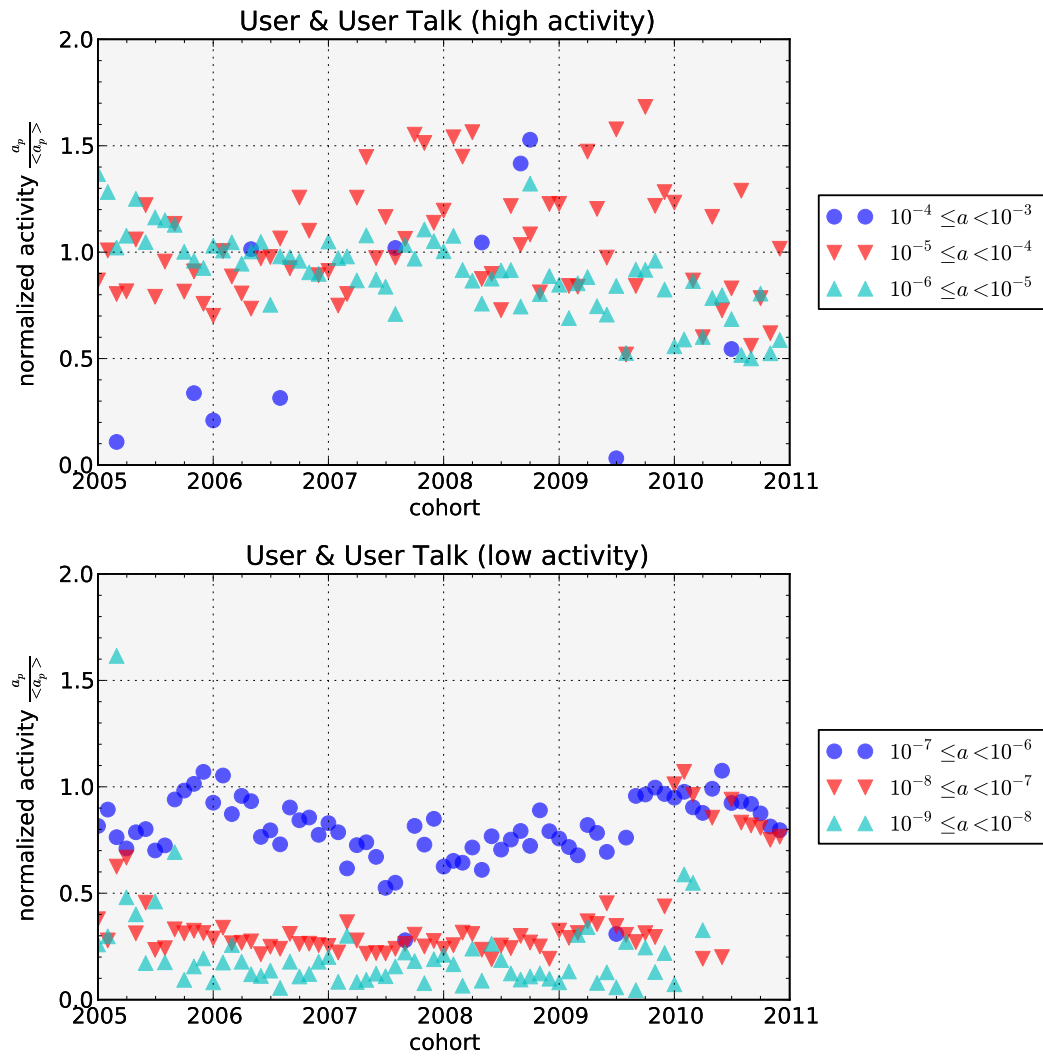


Figure 7.10. Normalized activity rate over time, edits to NS 2 & 3. Top: high activity cohorts. Bottom: low activity cohorts.

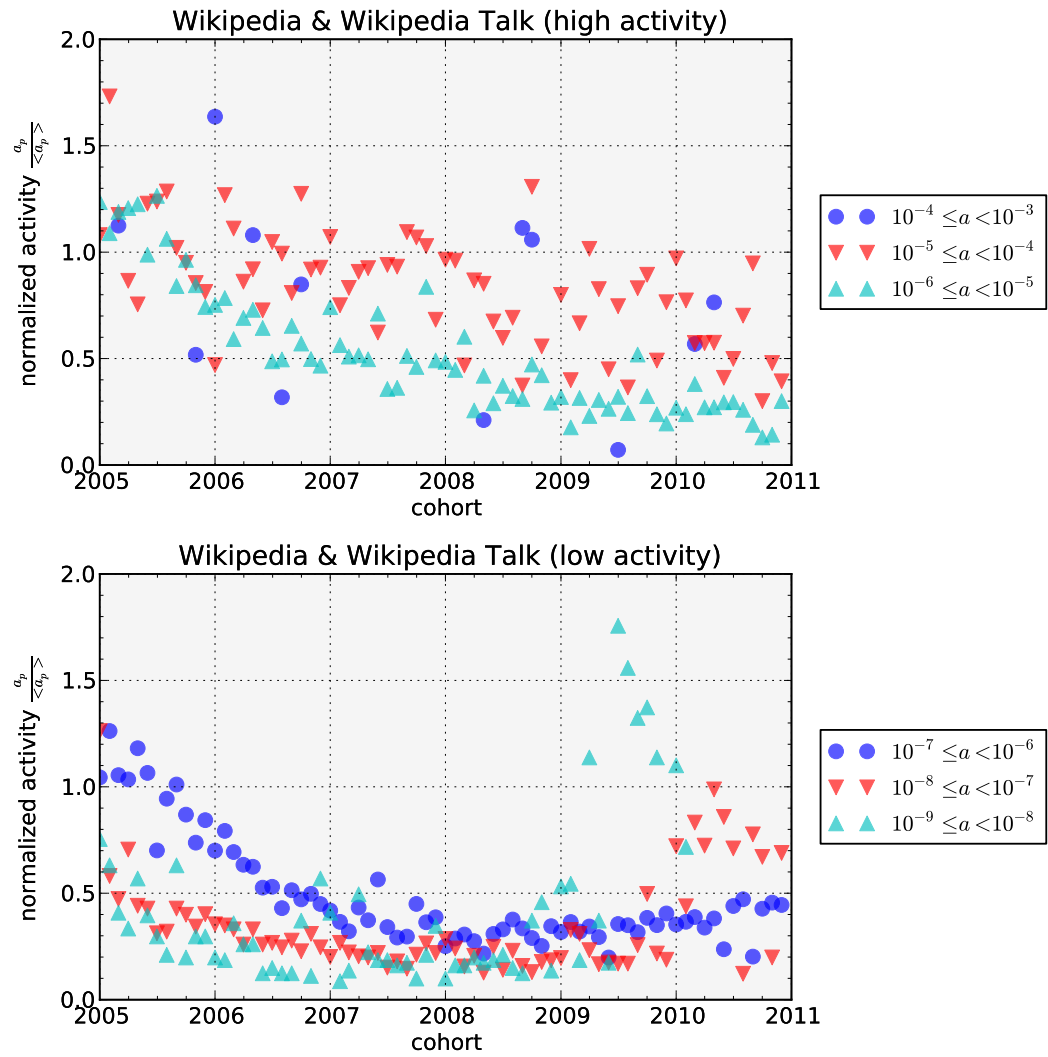


Figure 7.11. Normalized activity rate over time, edits to ns 4 & 5. Top: high activity cohorts. Bottom: low activity cohorts.

Chapter 8

Conclusions

One truth was that you got out of such models just about what you put into them. They were mostly a way of recording your preconceptions. The second truth was that such preconceptions could be fluently incorporated into models of this type and could, in fact, reach any conclusions that one wanted. Hence, at one level, this work could be construed as a criticism of all model-building in which one's preconceptions were not tested against data.

Leo Kadanoff

In this final chapter we critically evaluate the findings of this dissertation, and discuss a potential agenda for future research on user participation in peer production systems. We make explicit reference to the operational research questions enumerated in the introduction of the dissertation (see 1.3.2).

8.1 The distribution of the lifespan of user activity

In chapter 3 we analyzed the distribution of the lifespan of user activity in five of the largest Wikipedia communities, found strong support for the fact that user activity lifespan is distributed according to a mixture of multiple log-normal

distributions (QUESTION 1a), measured the mixture coefficients across all five communities (QUESTION 1b), and observed the temporal evolution of the mixture coefficients (QUESTION 1c). Finally, in chapter 7 we analyzed how user editing activity changes over the course of the lifespan of activity of users, and throughout the history of the community (QUESTION 1d).

Regarding all these research questions, a few considerations need to be taken into account. We start focusing on those related to the work of chapter 3 and separately address the last one below.

Even though the empirical distribution of user participation from Wikipedia shows a clearly heavy-tailed decay, we found that there is little or no support for a power-law hypothesis. Other distributions have been proposed to model the period of user participation in peer production systems, most notably the exponential distribution. We did not check explicitly whether there is support for an exponential decay in our data simply because our findings seem to be quite straightforward on the multi-modality of the lifespan of user accounts in Wikipedia.

In particular, our analysis found that the model that provides the best description for our data is a mixture of three log-normal components. Clearly, more analysis is needed in order to assess whether the data support this hypothesis. In particular, some limitations of the present analysis should be considered.

The first is that we did not take into account the rate at which users join Wikipedia. This information is important because the estimation of the overall distribution of user lifespan does depend on this quantity: since we only have access to an observation window of finite length, any large, sudden fluctuation in the number of new users might result in an excess of observed frequency for the corresponding values of the activity lifespan. In fact, it is well-known that Wikipedia underwent an exponential phase of growth during the 2006–2008 period. Without any information about the relative importance of each generation of new users, the mixture model we employ can only adapt to these “bumps” by introducing more components than necessary.

The second limitation of this kind of analysis is that the results of our estimation only apply to a specific stage of development of each community. This is a consequence of our choice of dealing with a truncated sample instead of a censored one. Censoring can be dealt with using very similar tools to those we use to perform our estimation, so an interesting development of this work would be to perform our fit using a censored model, and see whether the lifespan distribution is stable over time and whether a three components mixture is still the favored model.

8.1.1 Temporal trends of user activity

The longitudinal study of user activity we performed in chapter 7 sheds more light on the evolution of editing activity in the English Wikipedia during its development. In particular, we found that editor activity is strongly inhomogeneous over time. This finding might inform the modeling activity in the future, in particular allowing inhomogeneous activity in our peer production model and a meaningful distribution of user activity rates. Another direction of research would be an investigation of the mechanisms governing the evolution of the stretched exponential model of activity decay. An attempt to formulate a general model that could reconcile the observed patterns of editing activity for those cohorts marked by high rates of editing activity would be appealing as well. In particular, investigating whether the dynamics of collective attention (e.g. (Wu and Huberman [2007])) towards the Wikipedia project explains these phenomena would be especially interesting.

8.2 An agent-based approach to modeling peer production

In chapter 4 we introduced a microscopic, agent-based model of user participation in a peer production community (QUESTION 2a), studied its parameters by means of global sensitivity analysis in chapter 5 (QUESTION 2b & QUESTION 2c), and calibrated them by means of indirect inference in chapter 6 (QUESTION 2d).

One common complaint about the agent-based methodology is that what is often presented as the “main” behavior of the model is usually attained only for a specific range of values of its parameters. As Epstein and Axtell [1996] points out, being able to generate the expected behavior is the minimum one has to do in order to give credibility to the model one is proposing. Of course this not enough to claim that the rules agents follow in the simulations are actually the ones that in reality govern the collective behavior under investigation.

Often it is not clear whether certain collective patterns emerge because some assumptions were favored over others, and sometimes it is not even clear which parameters are the ones really responsible for such emergent behaviors.

Being able to rank parameters by the effect they have on the response of the model is thus very important, since it gives a tool to discern which parameters are really important and which are not, and that might be removed in order to yield a simpler model. It also helps to focus attention on fewer factors and thus opens up the possibility of proposing alternative models for comparison.

In the context of agent-based models of social group dynamics, our application, in chapter 5, of global sensitivity analysis is, to our knowledge, new. As more and more datasets about social interactions are produced, and better ways to quantitatively measure cultural traits are devised, we can foresee that these kind of models will increasingly find empirical application for the study of opinion and cultural formation and, in general, group dynamics. Global sensitivity analysis will be a useful tool to quantitatively assess the importance of various parameters and, hence, guide the modeler in the empirical investigation of social phenomena with agent-based models.

The results of the calibration (QUESTION 2d) performed in chapter 6 are encouraging but certainly not conclusive. In particular, a comparison with empirical data shows that indirect inference suffers from a certain upward bias in estimating both the confidence parameter ε and the initial number of successes c_s . Several steps could be taken to improve this situation. First, the model could be simplified, for example following the results of the factor screening analysis. This would result in a simplified simulation scenario. Ideally, running the indirect inference only on two parameters – ε and c_s – should give better results than on five. Alternative indirect inference techniques could be explored too. We could use different distance criteria than that given by (6.2), for example minimizing the likelihood of the auxiliary model. This approach is often called simulated quasi-maximum likelihood (or SQML). An entirely different calibration technique could be used as well, for example optimization by means of genetic algorithms.

Besides these enhancements, we should also consider the current limitations of the indirect inference technique. The most pressing is the inability to compute any measure of goodness of our model in fitting the data, since we do not have access to the likelihood of the data given the parameters estimated via indirect inference. This means, among other things, that we cannot (yet) perform any hypothesis testing with our peer production model. Again, this limitation could be easily overcome by adopting a SQML framework.

8.2.1 Beyond bounded confidence and other developments

Thus a natural development of this research will be to explore different hypotheses than those currently employed in our model of peer production. We have already discussed the idea of introducing a more realistic model for the temporal patterns of the editing of Wikipedia users; this work will of course be informed by the findings of chapter 7.

Another part of the model that could be further refined is the page selection

model. In fact, the factor screening analysis revealed that the popularity factor c_p has almost no effect in determining the average user lifespan (QUESTION 2b, see chapter 5), and similarly for the sufficient statistics of the Gaussian mixture model (chapter 6). On the other hand, allowing users to select pages using only global information might have an effect in determining the overall cultural dynamics of group formation – most notably the emergence of a mono-culture for high values of ε (see simulations in chapter 4). Instead, other mechanisms could be introduced to account for user interests. A self-reinforcing mechanism like the one employed by Crandall et al. [2008] could be employed for example.

Finally, a promising direction of research would be to explore different mechanisms of attitude change, and in particular opinion-averaging rules other than bounded confidence. Inspired by the recent results of Flache and Macy [2011] we could for example introduce noise in the interaction between user and page, and allow for group-level interactions instead of dyadic interactions, which would probably result in a more accurate description of the phenomenon of social influence that underlies our model of community formation.

8.3 Future works and generalizations of the present work

Based on the previous considerations, we can sketch a few directives for future works:

1. *Estimation of activity lifespan distribution with censoring.* As already stated, parameters estimation using censoring techniques should be performed in order to compare effectively different models, including the multimodal lognormal model that was proposed in this thesis.
2. *Influx of new users.* Likewise, an improved estimation methodology that takes into account the empirical rate of new user registrations should be developed, in order to understand how the influx of new users affects and potentially biases the measurements of the activity lifespan of users.
3. *Lifecycle of user activity.* The evolution of user activity rates should be investigated more deeply, in particular moving from purely empirical analysis to a modeling setting. It would be interesting to devise a model of user attention and motivation that could predict the observed patterns of activity decay as a function of the average productivity. In this way both the stretched exponential decay that we see at low activity cohorts and the

highly inhomogeneous patterns that we saw for high activity cohorts could be reconciled under the same parametric model.

4. *Improved model calibration and hypothesis testing.* As we said already in the previous section, an improved model calibration could be performed by simplifying the model and using different indirect inference techniques. In particular, the ability to compute p -value would open up the road to hypothesis testing studies and thus comparison of different hypotheses about user participation.
5. *General model improvements.* We refer the reader to the dedicated section about model limitation in Chapter 4.
6. *Alternative models of norm formation.* As evidenced in the previous section, different models of cultural formations, including noise, could be explored.
7. *Connection with other peer production systems.* The results presented in this work are about Wikipedia, a prominent peer production system. It would be interesting to see how our findings generalize to other peer productions systems, in particular Free/Open Source Software teams. The first step could be to gather data about user activity lifespan, and see if a multimodal Log-normal distribution also holds in these systems. The second step would be to see if our peer production model could be applied to this case. As we have said, social production norms are omnipresent in peer production communities, and thus it should not be difficult to motivate the usage of our model. However, in FOSS teams the role of intrinsic remuneration is more important than in Wikipedia, and thus this factor should be taken into consideration when proposing a generative explanation like the one we have in this thesis. Similarly it would also be interesting to expand the analysis to other crowdsourcing applications, such as the Mechanical Turk or other sharing websites such as Flickr, or Youtube.



This chapter opens with a quotation from physicist Leo Kadanoff who, looking back on his brief work on urban models of the early 60s, lamented that those kinds of models were just a way of recording the preconception of the modeler. Many things have happened in science since then, but we can surely affirm that many of the agent-based models that are proposed today in scientific literature are still affected by the same problem. The most important contribution of this

dissertation is to show that this need not be the case anymore. The standard toolkit of scientific investigation, which is based on the comparison against empirical data, can indeed be applied to the study of collective social phenomena with simulation models.

Bibliography

- Abbate, J. [1999]. *Inventing the Internet*, The MIT Press, Cambridge, MA, USA.
- Adar, E. and Huberman, B. A. [2000]. Free riding on Gnutella, *First Monday* 5(10). Last accessed: 2011-09-05 19:40:01.
- Adar, E., Zhang, L. and Lukose, R. M. [2004]. Implicit structure and the dynamics of blogspace, *WWW 2004 Workshop on the Weblogging Ecosystem : Aggregation, Analysis and Dynamics*.
- Adler, B. T. and Alfaro, L. D. [2007]. A content-driven reputation system for the wikipedia, *Proceedings of the 16th international conference on World Wide Web*, pp. 261–270.
- Adomavicius, G. and Tuzhilin, A. [2005]. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, *IEEE transactions on knowledge and data engineering* 17(6): 734–749.
- Aiello, L.-M., Barrat, A., Cattuto, C., Ruffo, G. and Schifanella, R. [2010]. Link creation and profile alignment in the anobii social network, *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pp. 249 –256.
- Albert, R. and Barabási, A.-L. [2002]. Statistical mechanics of complex networks, *Rev. Mod. Phys.* 74: 47–97.
- Alexa.com [2011]. Wikipedia.org site info. Last accessed 2011-09-14.
URL: <http://www.alexacom/siteinfo/wikipedia.org>
- Amrita, C. and van Hilleghersberg, J. [2008]. Detecting coordination problems in collaborative software development environments, *Information Systems Management* 25(1): 57–70.
- Anthony, D., Smith, S. W. and Williamson, T. [2009]. Reputation and reliability in collective goods: The case of the online encyclopedia wikipedia, *Rationality and Society* 21: 283–306.

- Axelrod, R. and Hamilton, W. [1981]. The evolution of cooperation, *Science* **211**(4489): 1390–1396.
URL: <http://www.sciencemag.org/content/211/4489/1390.abstract>
- Backstrom, L., Huttenlocher, D., Kleinberg, J. and Xiangyang, L. [2006]. Group formation in large social networks: Membership, growth, and evolution, *Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA, USA.
- Backstrom, L., Kumar, R., Marlow, C., Novak, J. and Tomkins, A. [2007]. Preferential behavior in online groups, *Proc. of WSDM'08*, pp. 1–11.
- Baldi, P., Frasca, P. and Smyth, P. [2003]. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, John Wiley & Sons.
- Ball, P. [2002]. The physical modelling of society: a historical perspective, *Physica A: Statistical Mechanics and its Applications* **314**(1–4): 1–14.
- Barabási, A.-L. and Albert, R. [1999]. Emergence of scaling in random networks, *Science* **286**(5439): 509–512.
- Barabási, A.-L. [2005]. The origin of bursts and heavy tails in human dynamics, *Nature* **435**(7039): 207–211.
- Bartholdi, J. J., Tovey, C. A. and Trick, M. A. [1989]. The computational difficulty of manipulating an election, *Social Choice and Welfare* **6**: 227–241. 10.1007/BF00295861.
- Ben-Naim, E. [2005]. Opinion dynamics: Rise and fall of political parties, *EPL (Europhysics Letters)* **69**(5): 671–677.
- Benkler, Y. [2002]. Coase's penguin, or Linux and the nature of the firm, *Yale Law Journal* **112**(3): 369–446.
- Bernardes, A., Stauffer, D. and Kertész, J. [2002]. Election results and the sznajd model on barabasi network, *The European Physical Journal B - Condensed Matter and Complex Systems* **25**: 123–127.
- Beschastnikh, I., Kriplean, T. and McDonald, D. W. [2008]. Wikipedian self-governance in action: Motivating the policy lens, *Proceedings of the second ICWSM conference*.
- Bianchi, C., Cirillo, P., Gallegati, M. and Vagliasindi, P. [2007]. Validating and calibrating agent-based models: A case study, *Computational Economics* **30**(3): 245–264.

- Bird, C., Pattison, D., D'Souza, R., Filkov, V. and Devanbu, P. [2008]. Latent social structure in open source projects, *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*, SIGSOFT '08/FSE-16, ACM, New York, NY, USA, pp. 24–35.
- Bishop, A. P and Starr, S. L. [1996]. Social informatics of digital library use and infrastructure, *Annual Review of Information Science and Technology* **31**: 301–401.
- Bishop, C. M. [2006]. *Pattern Recognition and Machine Learning*, Springer-Verlag NY, Secaucus, NJ, USA.
- Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D.-U. [2006]. Complex networks: Structure and dynamics, *Physics Reports* **424**(4-5): 175 – 308.
- Bonabeau, E. [2002]. Agent-based modeling: Methods and techniques for simulating human systems, *Proceedings of the National Academy of Sciences of the United States of America* **99**(Suppl 3): 7280–7287.
- Bowman, I. T. and Holt, R. C. [1998]. Software architecture recovery using conway's law, *Proceedings of the 1998 conference of the Centre for Advanced Studies on Collaborative research*, CASCON '98, IBM Press.
- Brandes, U., Kenis, P., Lerner, J. and van Raaji, D. [2009a]. Is editing more rewarding than discussion? a statistical framework to estimate causes of dropout from wikipedia, *Proceedings of WWW*.
- Brandes, U., Kenis, P., Lerner, J. and van Raaji, D. [2009b]. Network analysis of collaboration structure in wikipedia, *Proceedings of the 18th international conference on World wide web*.
- Brandes, U. and Lerner, J. [2007]. Revision and co-revision in Wikipedia., *Proc. Intl. Workshop Bridging the Gap Between Semantic Web and Web 2.0, 4th Europ. Semantic Web Conf. (ESWC'07)*.
- Brandes, U. and Lerner, J. [2008]. Visual analysis of controversy in user-generated encyclopedias, *Information Visualization* **7**: 34–48.
- Brewington, B. E. and Cybenko, G. [2000]. How dynamic is the web?, *Computer Networks* **33**(1-6): 257 – 276.
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. [2000]. Graph structure in the web, *Computer Networks* **33**(1-6): 309–320.

- Brooks, Jr., F. P. [1995]. *The mythical man-month (anniversary ed.)*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Bryant, S. L., Forte, A. and Bruckman, A. [2005]. Becoming wikipedia: transformation of participation in a collaborative online encyclopedia, *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, ACM, New York, NY, USA, pp. 1–10.
- Burke, M. and Kraut, R. [2008]. Mopping up: modeling wikipedia promotion decisions, *CSCW '08: Proceedings of the 2008 ACM conference on Computer supported cooperative work*, ACM, New York, NY, USA, pp. 27–36.
- Burke, R. [2002]. Hybrid recommender systems: Survey and experiments, *User Modeling and User-Adapted Interaction* **12**: 331–370. 10.1023/A:1021240730564.
- Byrd, R. H., Lu, P. and Nocedal, J. [1995]. A limited memory algorithm for bound constrained optimization, *SIAM Journal on Scientific and Statistical Computing* **16**(5): 1190–1208.
- Capocci, A., Rao, F. and Caldarelli, G. [2008]. Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia wikipedia, *EPL (Europhysics Letters)* **81**(2): 28006.
- Capocci, A., Servedio, V. D. P., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S. and Caldarelli, G. [2006]. Preferential attachment in the growth of social networks: The Internet encyclopedia Wikipedia, *Phys. Rev. E* **74**(3): 036116.
- Carletti, T., Fanelli, D., Grolli, S. and Guarino, A. [2006]. How to make an efficient propaganda, *EPL (Europhysics Letters)* **74**(2): 222–228.
- Carletti, T., Fanelli, D., Guarino, A., Bagnoli, F. and Guazzini, A. [2008]. Birth and death in a continuous opinion dynamics model, *Eur. Phys. J. B* **64**(2): 285–292.
- Carletti, T., Fanelli, D. and Righi, S. [2010]. On the evolution of a social network, *ArXiv e-prints*.
- Castellano, C. [2005]. Effect of network topology on the ordering dynamics of voter models, in J. Marro, P. L. Garrido and M. A. Munoz (eds), *AIP Conference Proceedings*, Vol. 779, AIP, pp. 114–120.
- Castellano, C., Fortunato, S. and Loreto, V. [2009]. Statistical physics of social dynamics, *Rev. Mod. Phys.* **81**(2): 591–646.

- Castellano, C., Vilone, D. and Vespignani, A. [2003]. Incomplete ordering of the voter model on small-world networks, *EPL (Europhysics Letters)* **63**(1): 153.
- Chan, K., Saltelli, A. and Tarantola, S. [2000]. Winding stairs: A sampling tool to compute sensitivity indices, *Statistics and Computing* **10**: 187–196.
- Chatterjee, S. and Seneta, E. [1977]. Towards consensus: Some convergence theorems on repeated averaging, *Journal of Applied Probability* **14**(1): 89–97.
- Chevaleyre, Y., Endriss, U., Lang, J. and Maudet, N. [2007]. A short introduction to computational social choice, in J. van Leeuwen, G. Italiano, W. van der Hoek, C. Meinel, H. Sack and F. Plášil (eds), *SOFSEM 2007: Theory and Practice of Computer Science*, Vol. 4362 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 51–69.
- Chmiel, A. and Hołyst, J. A. [2010]. Flow of emotional messages in artificial social networks, *International Journal of Modern Physics C* **21**: 593–602.
- Choi, B., Alexander, K., Kraut, R. E. and Levine, J. M. [2010]. Socialization tactics in wikipedia and their effects, *CSCW '10: Proceedings of the 2010 ACM conference on Computer supported cooperative work*, ACM, New York, NY, USA, pp. 107–116.
- Ciampaglia, G. L. [2011]. A bounded confidence approach to understand user participation in peer production systems, *Third International Conference on Social Informatics (SocInfo'11)*, LNCS, Springer Verlag, Singapore.
- Ciampaglia, G. L. and Vancheri, A. [2010]. Empirical analysis of user participation in online communities: the case of Wikipedia, *Proceedings of ICWSM*.
- Ciffolilli, A. [2003]. Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of wikipedia, *First Monday* **8**(12).
URL: <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1108/1028>
- Clauset, A., Shalizi, C. R. and Newman, M. E. J. [2009]. Power-law distributions in empirical data, *SIAM Review* **51**(4): 661–703.
- Clifford, P. and Sudbury, A. [1973]. A model for spatial conflict, *Biometrika* **60**(3): 581–588.
- Cockburn, A. [2001]. *Writing Effective Use Cases*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

- Cohen, J. E., Hajnal, J. and Newman, C. M. [1986]. Approaching consensus can be delicate when positions harden, *Stochastic Processes and their Applications* **22**(2): 315–322.
- Conway, M. [1968]. How do committees invent, *Datamation* **14**(5): 28–31.
- Cosley, D., Frankowski, D., Kiesler, S., Terveen, L. and Riedl, J. [2005]. How oversight improves member-maintained communities, *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, pp. 11–20.
- Cosley, D., Frankowski, D., Terveen, L. and Riedl, J. [2007]. Suggestbot: using intelligent task routing to help people find work in wikipedia, *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, ACM, Honolulu, Hawaii, USA, pp. 32–41.
- Costa, L. d. F., Rodrigues, F. A., Travieso, G. and Villas Boas, P. R. [2007]. Characterization of complex networks: A survey of measurements, *Advances in Physics* **56**(1): 167–242.
- Cox, J. T. and Griffeath, D. [1986]. Diffusive clustering in the two dimensional voter model, *The Annals of Probability* **14**(2): 347–370.
- Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. and Suri, S. [2008]. Feedback effects between similarity and social influence in online communities, *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- D'Agostino, R. B., Belanger, A. and D'Agostino, Ralph B., J. [1990]. A suggestion for using powerful and informative tests of normality, *The American Statistician* **44**(4): 316–321.
- Dancik, G. M., Jones, D. E. and Dorman, K. S. [2010]. Parameter estimation and sensitivity analysis in an agent-based model of leishmania major infection, *Journal of Theoretical Biology* **262**(3): 398 – 412.
- Daughety, A. and Reinganum, J. [1999]. Stampede to judgement: persuasive influence and herding behavior by courts, *Am Law Econ Rev* **1**(1): 158–189.
- Davis, J. H. [1973]. Group decision and social interaction: A theory of social decision schemes, *Psychological Review* **80**(2): 97 – 125.
- de Oliveira, M. J., Mendes, J. F. F. and Santos, M. A. [1993]. Nonequilibrium spin models with ising universal behaviour, *Journal of Physics A: Mathematical and General* **26**(10): 2317.

- de Solla Price, D. J. [1965]. Networks of scientific papers, *Science* **149**(3683): 510–515.
- Deffuant, G., Amblard, F., Weisbuch, G. and Faure, T. [2002]. How can extremism prevail? a study based on the relative agreement interaction model, *J. Art. Soc. Soc. Sim.* **5**(4): paper 1.
- Deffuant, G., Neau, D., Amblard, F. and Weisbuch, G. [2001]. Mixing beliefs among interacting agents, *Adv. Comp. Sys.* **3**: 87–98.
- DeGroot, M. H. [1974]. Reaching a consensus, *Journal of the American Statistical Association* **69**(345): 118–121.
- Dellarocas, C. [2003]. The digitization of word of mouth: Promise and challenges of online feedback mechanisms, *Manag. Sci.* **49**(10): 1407–1424.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. [1977]. Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society, B* **39**(1): 1–38.
- Denning, P., Horning, J., Parnas, D. and Weinstein, L. [2005]. Wikipedia risks, *Commun. ACM* **48**(12): 152–152.
- DeSanctis, G. and Fulk, J. (eds) [1999]. *Shaping organization form: communication, connection, and community*, Sage, Newsbury Park, CA.
- Dezsö, Z., Almaas, E., Lukács, A., Rácz, B. and Barabási, A.-L. [2006]. Dynamics of information access on the web, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **73**(6): 066132.
- Dorogovtsev, S. and Mendes, J. [2003]. *Evolution of networks: from biological nets to the Internet and WWW*, Oxford University Press, New York, NY, USA.
- Dorogovtsev, S. N. and Mendes, J. F. F. [2000]. Evolution of networks with aging of sites, *Phys. Rev. E* **62**: 1842–1845.
- Dorogovtsev, S. N., Mendes, J. F. F. and Samukhin, A. N. [2000]. Structure of growing networks with preferential linking, *Phys. Rev. Lett.* **85**: 4633–4636.
- Druck, G., Gerome, M. and McCallum, A. [2008]. Learning to predict the quality of contributions to wikipedia, *Proceedings of the second ICWSM conference*, pp. 7–12.
- Durrett, R. [2007]. *Random Graph Dynamics*, number 20 in *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press.

- Dutton, W. (ed.) [1997]. *Information and communication technologies: Vision & realities*, Oxford University Press, New York, NY, USA.
- Ebel, H., Mielsch, L.-I. and Bornholdt, S. [2002]. Scale-free topology of e-mail networks, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **66**(3): 035103.
- Ellner, S. P. and Guckenheimer, J. [2006]. *Dynamic Models in Biology*, Princeton University Press.
- Encyclopædia Britannica [2006]. Fatally flawed: refuting the recent study on encyclopedic accuracy by the journal Nature. Last accessed: 2011-09-12.
URL: http://corporate.britannica.com/britannica_nature_response.pdf
- Epstein, J. M. and Axtell, R. [1996]. *Growing artificial societies: social sciences from the bottom up*, The MIT Press, Cambridge, MA, USA.
- Erdős, P. and Rényi, A. [1959]. On random graphs, i, *Publicationes Mathematicæ* **6**: 290–297.
- Faloutsos, M., Faloutsos, P. and Faloutsos, C. [1999]. On power-law relationships of the internet topology, *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, SIGCOMM '99, ACM, New York, NY, USA, pp. 251–262.
- Festinger, L. [1950]. Informal social communication, *Psychological Review* **57**(5): 271–282.
- Festinger, L. and Thibaut, J. [1951]. Interpersonal communication in small groups., *J Abnorm Psychol.* **46**(1): 92–99.
- Filho, R. N. C., Almeida, M. P., Andrade, J. S. and Moreira, J. E. [1999]. Scaling behavior in a proportional voting process, *Phys. Rev. E* **60**: 1067–1068.
- Flache, A. and Macy, M. W. [2011]. Local convergence and global diversity: From interpersonal to social influence, *Journal of Conflict Resolution* pp. 1–26.
- Forte, A. and Bruckman, A. [2008]. Scaling consensus: Increasing decentralization in wikipedia governance, *Proc. of Hawaii International Conference on System Sciences*, Waikoloa, Big Island, Hawaii.
- Fortunato, S. [2004]. Universality of the threshold for complete consensus for the opinion dynamics of Deffuant et al., *International Journal of Modern Physics C* **15**: 1301–1307.

- Fortunato, S. [2010]. Community detection in graphs, *Physics Reports* **486**(3-5): 75 – 174.
- Fortunato, S. and Castellano, C. [2007]. Scaling and universality in proportional elections, *Phys. Rev. Lett.* **99**(13): 138701.
- Friedkin, N. E. [2001]. Norm formation in social influence networks, *Social Networks* **23**(3): 167–189.
- Friedkin, N. E. and Johnsen, E. C. [1990]. Social influence and opinions, *The Journal of Mathematical Sociology* **15**(3-4): 193–206.
- Friedkin, N. and Johnsen, E. [1999]. Social influence networks and opinion change, *Advances in Group Processes* p. 1–29.
- Fung, H., van Liere, D. and Moeller, E. [2011]. Editor trends study: Summary of findings, *Technical report*, Wikimedia Foundation, inc.
- Fürbinger, J. M. [1996]. Sensitivity analysis for modellers, *Air Infiltration Review* **17**(4): 8–10.
- Galam, S. [1986]. Majority rule, hierarchical structures, and democratic totalitarianism: A statistical approach, *Journal of Mathematical Psychology* **30**(4): 426–434.
- Galam, S. [2002]. Minority opinion spreading in random geometry, *The European Physical Journal B - Condensed Matter and Complex Systems* **25**: 403–406. 10.1140/epjb/e20020045.
- Galam, S. [2005]. Heterogeneous beliefs, segregation, and extremism in the making of public opinions, *Phys. Rev. E* **71**: 046123.
- Galam, S., Gefen, Y. and Shapir, Y. [1982]. Sociophysics: A new approach of sociological collective behaviour. i. meanbehaviour description of a strike, *J. Math. Sociol.* **9**(1): 1–13.
- Galton, F. [1907]. Vox populi, *Nature* **75**: 450–451.
- Gil, S. and Zanette, D. H. [2006]. Coevolution of agents and networks: Opinion spreading and community disconnection, *Physics Letters A* **356**(2): 89 – 94.
- Giles, J. [2005]. Internet encyclopaedias go head to head, *Nature* **438**(7070): 900–901.
- Gillespie, D. [1977]. Exact stochastic simulation of coupled chemical reactions, *Journal of Physical Chemistry* **81**(25): 2340–2361.

- Gilliland, S. W., Benson, L. I. and Schepers, D. [1998]. A rejection threshold in justice evaluation: effects on judgement and decision making, *Organizational behavior and Human decision process* **76**: 113–131.
- Golbeck, J. A. [2005]. *Computing and applying trust in web-based social networks*, PhD thesis, University of Maryland at College Park College Park.
- Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. [1992]. Using collaborative filtering to weave an information tapestry, *Commun. ACM* **35**: 61–70.
- Golder, S. A. and Huberman, B. A. [2006]. Usage patterns of collaborative tagging systems, *Journal of Information Science* **32**(2): 198–208.
- Goldman, E. [2009]. Wikipedia’s labor squeeze and its consequences, *Telecomm. and High Tech. Law* **8**: 157–184.
- Gonzalez-Bailon, S., Banchs, R. E. and Kaltenbrunner, A. [2010]. Emotional reactions and the pulse of public opinion: Measuring the impact of political events on the sentiment of online discussions, *ArXiv e-prints* .
- Gonçalves, B. and Ramasco, J. [2008]. Human dynamics revealed through web analytics, *Physical Review E* **78**(2): 026123.
- Gouriéroux, C. and Monfort, A. [1996]. *Simulation-Based Econometric Methods*, OUP/CORE Lecture Series, Oxford University Press, USA.
- Gouriéroux, C., Monfort, A. and Renault, E. [1993]. Indirect inference, *Journal Of Applied Econometrics* **8**(Suppl. S): 85–118. Conference On Econometric Inference Using Simulation Techniques, Rotterdam, Netherlands, Jun 05-06, 1992.
- Grabowski, A. and Kosiński, R. A. [2010]. Life span in online communities, *Phys. Rev. E* **82**(6): 066108.
- Gromov, G. [2000]. Roads and crossroads of the Internet history. Last accessed: 2011-09-05 11:44:41.
URL: http://www.netvalley.com/cgi-bin/intval/net_history.pl?chapter=2
- Guo, L., Tan, E., Chen, S., Zhang, X. and Zhao, Y. E. [2009]. Analyzing patterns of user content generation in online social networks, *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, pp. 369–378.

- Haken, H. [1978]. *Synergetics: an introduction. Non-equilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology*, Springer-Verlag, Berlin.
- Harder, U. and Paczuski, M. [2006]. Correlated dynamics in human printing behavior, *Physica A: Statistical Mechanics and its Applications* **361**(1): 329 – 336.
- Hardin, G. [1968]. The tragedy of the commons, *Science* **162**(3859): 1243–1248.
- Hars, A. and Ou, S. [2001]. Working for free? - motivations of participating in open source projects, *Hawaii International Conference on System Sciences* 7: 7014.
- Hegselmann, R. and Krause, U. [2002]. Opinion dynamics and bounded confidence–models, analysis, and simulation, *J. Art. Soc. Soc. Sim.* **5**(3): paper 2.
- Hegselmann, R. and Krause, U. [2006]. Truth and cognitive division of labour: First steps towards a computer aided social epistemology, *Journal of Artificial Societies and Social Simulation* **9**(3): 10.
- Herbsleb, J. D. and Grinter, R. E. [1999a]. Architectures, coordination, and distance: Conway’s law and beyond, *IEEE Software* **16**: 63–70.
- Herbsleb, J. D. and Grinter, R. E. [1999b]. Splitting the organization and integrating the code: Conway’s law revisited, *Proceedings of the 21st international conference on Software engineering, ICSE ’99*, ACM, New York, NY, USA, pp. 85–95.
- Herlocker, J. L., Konstan, J. A., Terveen, L. G. and Riedl, J. T. [2004]. Evaluating collaborative filtering recommender systems, *ACM T. Inf. Sys.* **22**(1): 5–53.
- Hill, W., Stead, L., Rosenstein, M. and Furnas, G. [1995]. Recommending and evaluating choices in a virtual community of use, *Proceedings of the SIGCHI conference on Human factors in computing systems, CHI ’95*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp. 194–201.
- Hill, W. and Terveen, L. [1996]. Using frequency-of-mention in public conversations for social filtering, *Proceedings of the 1996 ACM conference on Computer supported cooperative work, CSCW ’96*, ACM, New York, NY, USA, pp. 106–112.

- Hogg, T. and Huberman, B. A. [2008]. Solving the organizational free riding problem with social networks, *AAAI Spring Symposium on Social Information Processing*.
- Hogg, T. and Lerman, K. [2009]. Stochastic models of user-contributory web sites, *Proceedings of the Third International ICWSM Conference*, pp. 50–57.
- Hogg, T. and Szabo, G. [2009]. Diversity of user activity and content quality in online communities, *Proceedings of the Third International ICWSM Conference*.
- Holley, R. A. and Liggett, T. M. [1975]. Ergodic theorems for weakly interacting infinite systems and the voter model, *The Annals of Probability* **3**(4): 643–663.
- Holme, P. and Newman, M. E. J. [2006]. Nonequilibrium phase transition in the coevolution of networks and opinions, *Phys. Rev. E* **74**(5): 056108.
- Holyst, J., Kacperski, K. and Schweitzer, F. [2000]. Phase transitions in social impact models of opinion formation, *Physica A (Amsterdam, Neth.)* **285**: 199–210.
- Hu, M., Lim, E.-P. and Ramayya, K. [2009]. Predicting outcome for collaborative featured article nomination in wikipedia, *Proceedings of the Third International ICWSM Conference*.
- Huberman, B. A. and Adamic, L. [1999]. Growth dynamics of the world wide web, *Nature* **438**: 900–901.
- Huberman, B. A. and Hogg, T. [1995]. Communities of practice: Performance and evolution, *Computational & Mathematical Organization Theory* **1**: 73–92. 10.1007/BF01307829.
- Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E. and Lukose, R. M. [1998]. Strong regularities in world wide web surfing, *Science* **280**(5360): 95–97.
- Huff, C. and Finholt, T. (eds) [1994]. *Social issues in computing: Putting computing in its place*, McGraw-Hill, New York.
- Hui, P. and Buchegger, S. [2009]. Groupthink and peer pressure: Social influence in online social network groups, *Social Network Analysis and Mining, International Conference on Advances in* **0**: 53–59.
- Iñiguez, G., Kertész, J., Kaski, K. K. and Barrio, R. A. [2009]. Opinion and community formation in coevolving networks, *Phys. Rev. E* **80**: 066119.

- Iacono, C. S. and Weisband, S. [1997]. Developing trust in virtual teams, *Proceedings of the Hawaii International Conference on Systems Sciences*, Hawaii.
- Jansen, M., Rossing, W. and Daamen, R. [1994]. Monte-Carlo estimation of uncertainty contributions from several independent multivariate sources, in J. Grasman and G. van Straten (eds), *Predictability And Nonlinear Modelling In Natural Sciences And Economics*, pp. 334–343.
- Javanmardi, S., Lopes, C. and Baldi, P. [2010]. Modeling user reputation in wikis, *Statistical Analysis and Data Mining* 3(2): 126–139.
- Johansen, A. [2004]. Probing human response times, *Physica A (Amsterdam, Neth.)* 338(1-2): 286 – 291.
- Johansen, A. and Sornette, D. [2000]. Download relaxations dynamics on the WWW following newspaper publication of URL, *Physica A (Amsterdam, Neth.)* 276: 338–345.
- Johnson, N. L., Kotz, S. and Balakrishnan, N. [1994]. *Continuous Univariate Distributions*, Vol. 1 of *Wiley's Series in Probability and Statistics*, Wiley-Interscience.
- Kendall, B., Ellner, S., McCauley, E., Wood, S., Briggs, C., Murdoch, W. and Turchin, P. [2005]. Population cycles in the pine looper moth: Dynamical tests of mechanistic hypotheses, *Ecological Monographs* 75(2): 259–276.
- Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., Jaskolski, M. and Baker, D. [2011]. Crystal structure of a monomeric retroviral protease solved by protein folding game players, *Nat Struct Mol Biol* **On line version**: 1–3.
- Kiesler, S. (ed.) [1997]. *The culture of the Internet*, Lawrence Erlbaum Associates, Mahwah, NJ.
- Kittur, A., Chi, E., Pendleton, B., Suh, B. and Mytkowicz, T. [2007]. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie, *Alt. CHI*.
- Kittur, A., Pendleton, B. and Kraut, R. E. [2009]. Herding the cats: the influence of groups in coordinating peer production, *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (Wikysym '09)*, pp. 1–9.
- Kittur, A., Suh, B., Pendleton, B. A. and Chi, E. H. [2007]. He says, she says: conflict and coordination in wikipedia, *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 453–462.

- Kling, R. [2007]. What is social informatics and why does it matter?, *The Information Society* **23**(4): 205–220.
- Kling, R., Rosenbaum, H. and Hert, C. [1998]. Social informatics in information science: An introduction, *Journal of the American Society for Information Science* **49**(12): 1047–1052.
- Kollock, P. [1999]. *Communities in Cyberspace*, Routledge, London, chapter The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace, p. 220–239.
- Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R. and Riedl, J. [1997]. Grouplens: Applying collaborative filtering to usenet news, *Comm. ACM* **40**(3): 77–87.
- Kozma, B. and Barrat, A. [2008a]. Consensus formation on adaptive networks, *Phys. Rev. E* **77**(1): 016102.
- Kozma, B. and Barrat, A. [2008b]. Consensus formation on coevolving networks: groups' formation and structure, *Journal of Physics A: Mathematical and Theoretical* **41**(22): 224020.
- Krapivsky, P. L., Redner, S. and Leyvraz, F. [2000]. Connectivity of growing random networks, *Phys. Rev. Lett.* **85**: 4629–4632.
- Kriplean, T., Beschastnikh, I., McDonald, D. W. and Golder, S. A. [2007]. Community, consensus, coercion, control: Cs*w or how policy mediates mass participation, *Proceedings of the 2007 international ACM SIGGROUP conference on Supporting group work*, Sanibel Island, Florida, USA.
- Kumpula, J., Onnela, J.-P., Saramäki, J., Kertész, J. and Kaski, K. [2009]. Model of community emergence in weighted social networks, *Computer Physics Communications* **180**(4): 517 – 522.
- Laguna, M. F., Abramson, G. and Zanette, D. H. [2004]. Minorities in a model for opinion formation, *Complexity* **9**(4): 31–36.
- Lam, S. K. and Riedl, J. [2009]. Is wikipedia growing a longer tail?, *Proceedings of the ACM 2009 international conference on Supporting group work*, pp. 105–114.
- Lam, S., Karim, J. and Riedl, J. [2010]. The effects of group composition on decision quality in a social production community, *Proceedings of GROUP 2010*, Sanibel Island, Florida.

- Lamb, R. [1996]. Informational imperatives and socially mediated relationships, *The Information Society* **12**(1): 17–37.
- Lambiotte, R. [2007]. Activity ageing in growing networks, *Journal of Statistical Mechanics: Theory and Experiment* **2007**(02): 2–9.
- Lampe, C., Wash, R., Velasquez, A. and Ozkaya, E. [2010]. Motivations to participate in online communities, *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, ACM, New York, NY, USA, pp. 1927–1936.
- Latané, B. [1981]. The psychology of social impact, *American Psychologist* **36**(4): 343–356.
- Latané, B. [1996]. Dynamic social impact: The creation of culture by communication, *Journal of Communication* **46**(4): 13–25.
- Laughlin, P. [1980]. *Progress in Social Psychology*, Erlbaum, Hillsdale, NJ, chapter Social combination processes of cooperative problem-solving groups on verbal intellectual tasks, p. 127–155.
- Lazarsfeld, P. F. and Merton, R. K. [1954]. *Freedom and Control in Modern Society*, Van Nostrand, chapter Friendship as Social Process: a Substantive and Methodological Analysis.
- Lehmann, S., Jackson, A. D. and Lautrup, B. [2005]. Life, death and preferential attachment, *Europhys. Lett.* **69**: 298–303.
- Lerman, K. [2007a]. Social information processing in news aggregation, *IEEE Internet Computing* **11**(16): 16–28.
- Lerman, K. [2007b]. User participation in social media: Digg study, *Proc. of 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pp. 1–4.
- Lerner, J. and Tirole, J. [2002]. Some simple economics of open source, *The Journal of Industrial Economics* **50**(2): 197–234.
- Leskovec, J., Backstrom, L., Kumar, R. and Tomkins, A. [2008]. Microscopic evolution of social networks, *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, pp. 462–470.

- Leskovec, J., Huttenlocher, D. and Kleinberg, J. [2010]. Governance in social media: A case study of the wikipedia promotion process, *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'10)*.
- Lewenstein, M., Nowak, A. and Latané, B. [1992]. Statistical mechanics of social impact, *Phys. Rev. A* **45**: 763–776.
- Li, X. and Hitt, L. M. [2008]. Self-selection and information role of online product reviews, *Journal of Information systems research* **19**(4): 456–474.
- Liggett, T. M. [1985]. *Interacting Particle Systems*, Springer-Verlag, New York, USA.
- Lih, A. [2004]. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource, *5th International Symposium on Online Journalism*, University of Texas at Austin.
- Linstone, H. A. and Turoff, M. (eds) [1975]. *The Delphi Method : techniques and applications*, Addison-Wesley, Reading, MA, USA.
- Liu, J. and Ram, S. [2010]. Who does what: Collaboration patterns in the wikipedia and their impact on data quality, *19th Workshop on Information Technologies and Systems*.
- Lorenz, J. [2007a]. Continuous opinion dynamics under bounded confidence: A survey, *Intl J. Mod. Phys. C* **18**: 1819–1838.
- Lorenz, J. [2007b]. *Repeated Averaging and Bounded Confidence—Modeling, Analysis and Simulation of Continuous Opinion Dynamics*, PhD thesis, Universität Bremen.
- Lorenz, J. [2009]. Universality in movie rating distributions, *Eur. Phys. J. B* **71**(2): 251–258.
- Lorenz, J., Rauhut, H., Schweitzer, F. and Helbing, D. [2011]. How social influence can undermine the wisdom of crowd effect, *Proceedings of the National Academy of Sciences*.
- Lorenz, J. and Urbig, D. [2007]. About the power to enforce and prevent consensus by manipulating communication rules, *Adv. Comp. Sys.* **10**(2): 251–269.
- Malmgren, R. D., Hofman, J. M., Amaral, L. A. and Watts, D. J. [2009]. Characterizing individual communication patterns, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, pp. 607–616.

- Malmgren, R. D., Stouffer, D. B., Motterb, A. E. and Amaral, L. A. N. [2008]. A poissonian explanation for heavy tails in e-mail communication, *Proc. Natl. Acad. Sci. U. S. A.* **105**(47): 18153–18158.
- Martins, A. C. R. [2009]. Bayesian updating rules in continuous opinion dynamics models, *Journal of Statistical Mechanics: Theory and Experiment* **2009**(02): P02017.
- Maslow, A. H. [1954]. *Motivation and personality*, Harpers, Oxford, England.
- Mason, W. and Suri, S. [2011]. Conducting behavioral research on amazon's mechanical turk, *Behavior Research Methods* pp. 1–23. 10.3758/s13428-011-0124-6.
- McFadden, D. [1989]. A method of simulated moments for estimation of discrete response models without numerical integration, *Econometrica* **57**(5): 995–1026.
- McKay, M. D. [1992]. Latin hypercube sampling as a tool in uncertainty analysis of computer models, *Proceedings of the 24th conference on Winter simulation, WSC '92*, ACM, New York, NY, USA, pp. 557–564.
- McKenna, K. Y. A. and Green, A. S. [2002]. Virtual group dynamics., *Group Dynamics: Theory, Research, and Practice*. **6**(1): 116–127.
- McLachlan, G. J. and Jones, P. N. [1988]. Fitting mixture models to grouped and truncated data via the em algorithms, *Biometrics* **44**(2): 571–578.
- McPherson, M., Smith-Lovin, L. and Cook, J. M. [2001]. Birds of a feather: Homophily in social networks, *Annual Review of Sociology* **27**(1): 415–444.
- Michard, Q. and Bouchaud, J.-P. [2005]. Theory of collective opinion shifts: from smooth trends to abrupt swings, *Eur. Phys. J. B* **47**(1): 151–159.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. [2002]. Network motifs: Simple building blocks of complex networks, *Science* **298**(5594): 824–827.
- Moore, C. and Newman, M. E. J. [2000]. Epidemics and percolation in small-world networks, *Phys. Rev. E* **61**: 5678–5682.
- Mäs, M., Flache, A. and Helbing, D. [2010]. Individualization as driving force of clustering phenomena in humans, *PLoS Computational Biology* **6**(10): e1000959.

- Nagappan, N., Murphy, B. and Basili, V. [2008]. The influence of organizational structure on software quality: an empirical case study, *Proceedings of the 30th international conference on Software engineering*, ICSE '08, ACM, New York, NY, USA, pp. 521–530.
- Nardini, C., Kozma, B. and Barrat, A. [2008]. Who's talking first? consensus or lack thereof in coevolving opinion formation models, *Phys. Rev. Lett.* **100**: 158701.
- Neal, R. M. and Hinton, G. E. [1998]. *Learning in Graphical Models*, Kluwer Academic Publishers, Dordrecht, chapter A view of the EM algorithm that justifies incremental, sparse, and other variants, pp. 355–368.
- Nelder, J. A. and Mead, R. [1965]. A simplex method for function minimization, *The Computer Journal* **7**(4): 308–313.
- Newman, M. E. J. [2001]. Clustering and preferential attachment in growing networks, *Phys. Rev. E* **64**: 025102.
- Newman, M. E. J. [2002]. Spread of epidemic disease on networks, *Phys. Rev. E* **66**: 016128.
- Newman, M. E. J. [2003a]. Mixing patterns in networks, *Phys. Rev. E* **67**(2): 026126.
- Newman, M. E. J. [2003b]. The structure and function of complex networks, *SIAM Review* **45**(2): 167–256.
- Newman, M. E. J. and Girvan, M. [2004]. Finding and evaluating community structure in networks, *Phys. Rev. E* **69**: 026113.
- Newman, M. E. J., Watts, D. J. and Strogatz, S. H. [2002]. Random graph models of social networks, *Proceedings of the National Academy of Sciences of the United States of America* **99**(Suppl 1): 2566–2572.
- Nov, O. [2007]. What motivates wikipedians?, *Commun. ACM* **50**: 60–64.
- Nov, O., Naaman, M. and Ye., C. [2010]. Analysis of participation in an on-line photo-sharing community: A multidimensional perspective, *Journal of the American Society for Information Science and Technology (JASIST)* **61**(3): 1–12.
- Oliveira, J. G. and Barabási, A.-L. [2005]. Human dynamics: Darwin and einstein correspondence patterns, *Nature* **437**(7063): 1251.

- O'Reilly, T. [2005]. What is web 2.0—design patterns and business models for the next generation of software. Last accessed: 2011-09-13.
URL: <http://oreilly.com/web2/archive/what-is-web-20.html>
- Ortega, F. and Gonzales-Barahona, J. M. [2007]. Quantitative analysis of the wikipedia community of users, *Proceedings of WikiSym '07, 3rd Intl Symposium on Wikis*, Montréal, Québec, Canada.
- Ortega, F. and Izquierdo-Cortazar, D. [2009]. Survival analysis in open development projects, *Proceedings of the 2009 ICSE Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development*, FLOSS '09, IEEE Computer Society, Washington, DC, USA, pp. 7–12.
- Ostrom, E. [1990]. *Governing the Commons: The Evolution of Institutions for Collective Action*, Political Economy of Institutions and Decisions, Cambridge University Press, Cambridge, MA, USA.
- Ottaviani, M; Sørensen, P. [2001]. Information aggregation in debate: Who should speak first?, *Journal of Public Economics* **81**(3): 393–421.
- Palla, G., Barabási, A.-L. and Vicsek, T. [2007]. Quantifying social group evolution, *Nature* **446**(7136): 664–667.
- Panciera, K., Halfaker, A. and Terveen, L. [2009]. Wikipedians are born, not made, *Proceedings of GROUP'09*.
- Parameswaran, M. and Whinston, A. B. [2007]. Research issues in social computing, *Journal of the Association for Information Systems* **8**(6).
- Pemantle, R. [2007]. A survey of random processes with reinforcement, *Probability Surveys* **4**: 1–79.
- Plous, S. [1993]. *The psychology of judgement and decision making*, McGraw-Hill, Maidenhead, Berkshire, UK.
- Postmes, T., Spears, R., Sakhel, K. and de Groot, D. [2001]. Social influence in computer-mediated communication: The effects of anonymity on group behavior, *Pers Soc Psychol Bull* **27**(10): 1243–1254.
- Potthast, M., Stein, B. and Gerling, R. [2008]. Automatic vandalism detection in wikipedia, *Advances in Information Retrieval: 32nd European Conference on IR Research*, number 5993 in LNCS, Springer, pp. 663–668.

- Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L. and Riedl, J. [2007]. Creating, destroying and restoring value in wikipedia, *Proceedings of the 2007 international ACM SIGGROUP conference on Supporting group work*, Sanibel Island, Florida, USA.
- Quinn, A. J. and Bederson, B. B. [2011]. Human computation: a survey and taxonomy of a growing field, *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, ACM, New York, NY, USA, pp. 1403–1412.
- Raafat, R. M., Chater, N. and Frith, C. [2009]. Herding in humans, *Trends in Cognitive Sciences* **13**(10): 420 – 428.
- Raban, D. R., Moldovan, M. and Jones, Q. [2010]. An empirical study of critical mass and online community survival, *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, ACM, New York, NY, USA, pp. 71–80.
- Radicchi, F. [2009]. Human activity in the web, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **80**(2): 026118.
- Radicchi, F., Fortunato, S. and Castellano, C. [2008]. Universality of citation distributions: towards an objective measure of scientific impact, *Proc. Natl. Acad. Sci. USA* **105**,: 17268–17272.
- Rafaeli, S. and Ariel, Y. [2008]. *Psychological aspects of cyberspace: Theory, research, applications*, Cambridge University Press, chapter Online Motivational Factors: Incentives for Participation and Contribution in Wikipedia, pp. 243–267.
- Rainie, L. and Purcell, K. [2011]. How the public perceives community information systems, *Technical report*, Pew Internet & American Life Project. Last accessed: September 09, 2011.
URL: <http://www.pewinternet.org/Reports/2011/08-Community-Information-Systems.aspx>
- Raymond, E. S. [1999]. *The Cathedral and the Bazaar—Musings on Linux and Open Source by an accidental revolutionary*, O'Reilly & Associates, Inc., Sebastopol, CA.
- Reagle, J. M. J. [2007a]. Do as i do: Authorial leadership in wikipedia, *Proceedings of the 2007 international symposium on Wikis*, WikiSym '07, ACM, New York, NY, USA, pp. 143–156.

- Reagle, J. M. J. [2007b]. *Good Faith Collaboration—The Culture of Wikipedia*, The MIT Press.
- Ren, Y., Kraut, R. and Kiesler, S. [2007]. Applying common identity and bond theory to design of online communities, *Organization Studies* **28**(3): 377–408.
- Rescorla, E. [2001]. *SSL and TLS: Designing and Building Secure Systems*, Addison-Wesley Pub Co., United States.
- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J. [1994]. GroupLens: an open architecture for collaborative filtering of netnews, *Proceedings of the 1994 ACM conference on Computer supported cooperative work, CSCW '94*, ACM, New York, NY, USA, pp. 175–186.
- Resnick, P., Kuwabara, K., Zeckhauser, R. and Friedman, E. [2000]. Reputation systems, *Comm. ACM* **43**(12): 45–48.
- Robey, D. [1997]. *Steps to the future: fresh thinking on the dynamics of organizational transformation*, Jossey-Bass, San Francisco, CA, chapter The paradox of transformation: using contradictory logic to manage the organizational consequences of information technology.
- Roth, C. and Bourguine, P. [2005]. Epistemic communities: Description and hierarchical categorization, *Mathematical Population Studies* **12**(2): 107–130.
- Roth, C., Taraborelli, D. and Gilbert, N. [2008]. Measuring wiki viability – an empirical assessment of the social dynamics of a large sample of wikis, *Proceedings of the 4th International Symposium on Wikis (WikiSym '08)*.
- Sabater, J. and Sierra, C. [2005]. Review on computational trust and reputation models, *Art. Int. Rev.* **24**(1): 33–60.
- Salganik, M. J., Dodds, P. S. and Watts, D. J. [2006]. Experimental study of inequality and unpredictability in an artificial cultural market, *Science* **311**: 854–856.
- Salganik, M. J. and Watts, D. J. [2008]. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market, *Social Psychology Quarterly* **71**: 338–355.
- Salganik, M. J. and Watts, D. J. [2009]. Web-based experiments for the study of collective social dynamics in cultural markets, *Topics in Cognitive Science* **1**: 439–468.

- Saltelli, A., Tarantola, S., Campolongo, F. and Ratto, M. [2004]. *Sensitivity Analysis in Practice—A guide to Assessing Scientific Models*, John Wiley & Sons, Ltd.
- Salzarulo, L. [2006]. A continuous opinion dynamics model based on the principle of meta-contrast, *Journal of Artificial Societies and Social Simulation* **9**(1): 13.
- Sanger, L. M. [2009]. The fate of expertise after Wikipedia, *Episteme* **6**(1): 52–73.
- Santner, T., Williams, B. and Notz, W. [2003]. *The Design and Analysis of Computer Experiments*, Springer-Verlag, NY.
- Schaller, R. R. [1997]. Moore’s law: past, present, and future, *IEEE Spectrum* **34**: 52–59.
- Schelling, T. C. [1971]. Dynamic models of segregation, *The Journal of Mathematical Sociology* **1**(2): 143–186.
- Scheucher, M. and Spohn, H. [1988]. A soluble kinetic model for spinodal decomposition, *Journal of Statistical Physics* **53**: 279–294. 10.1007/BF01011557.
- Schroer, J. and Hertel, G. [2009]. Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it, *Media Psychology* **12**(1): 96–120.
- Schuler, D. [1994]. Social computing, *Commun. ACM* **37**: 28–29.
- Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C. and Leskove, J. [2008]. Mobile call graphs: beyond power-law and lognormal distributions, *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ’08)*, ACM, New York, NY, USA, pp. 596–604.
- Shalizi, C. R. [2008]. Social media as windows on the social life of the mind, *Proceedings of the 2008 AAAI symposium on “Social Information Processing”*.
- Shardanand, U. and Maes, P. [1995]. Social information filtering: algorithms for automating “word of mouth”, *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI ’95, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp. 210–217.
- Shatz, D. [2004]. *Peer review: a critical inquiry*, Issues in Academic Ethics, Rowman & Littlefield Publishers, Inc., Lanham, MD, USA.

- Sherif, M. [1936]. *The Psychology of Social Norms*, Harper, New Yor.
- Sherif, Muzafer; Hovland, C. I. [1961]. *Social judgment: Assimilation and contrast effects in communication and attitude change*, Yale University Press, Oxford, England.
- Slattery, S. P. [2009]. "edit this page": the socio-technological infrastructure of a wikipedia article, *SIGDOC '09: Proceedings of the 27th ACM international conference on Design of communication*, ACM, New York, NY, USA, pp. 289–296.
- Smith, A. A. [1993]. Estimating nonlinear time-series models using simulated vector autoregressions, *Journal of Applied Econometrics* **8**(S1): S63–S84.
- Smith, A. A. J. [2008]. indirect inference, in S. N. Durlauf and L. E. Blume (eds), *The New Palgrave Dictionary of Economics*, Palgrave Macmillan, Basingstoke.
- Smith, M. and Kollock, P. (eds) [1998]. *Communities in Cyberspace*, Routledge, London.
- Snow, R., O'Connor, B., Jurafsky, D. and Ng, A. Y. [2008]. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 254–263.
- Sobkowicz, P. [2009]. Modelling opinion formation with physics tools: Call for closer link with reality, *Journal Artificial Societies and Social Simulation* **12**(1): 11.
- Sobol', I. M. [2001]. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates, *Mathematics and Computers in Simulation* **55**(1-3): 271–280.
- Song, J. and Kim, Y. [2006]. Social influence process in the acceptance of a virtual community service, *Information Systems Frontiers* **8**: 241–252. 10.1007/s10796-006-8782-0.
- Sood, V. and Redner, S. [2005]. Voter model on heterogeneous graphs, *Phys. Rev. Lett.* **94**: 178701.
- Sornette, D. [2004]. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-organization and Disorder: Concepts and Tools*, Springer, Berlin / Heidelberg.

- Spier, R. [2002]. The history of the peer-review process, *Trends in Biotechnology* **20**(8): 357 – 358.
- Spinellis, D. and Louridas, P. [2008]. The collaborative organization of knowledge, *Commun. ACM* **51**(8): 68–73.
- Stasser, G., Kerr, N. and Davis, J. [1989]. *Psychology of Group Influence*, Lawrence Erlbaum, Hillsdale, NJ, chapter Influence processes and consensus models in decision-making groups, p. 279–326.
- Stauffer, D., Hohnisch, M. and Pittnauer, S. [2006]. The coevolution of individual economic characteristics and socioeconomic networks, *Physica A: Statistical Mechanics and its Applications* **370**(2): 734 – 740.
- Stauffer, D., Sousa, A. and Schulze, C. [2004]. Discretized opinion dynamics of the deffuant model on scale-free networks, *J. Art. Soc. Soc. Sim.* **7**(3): paper 7.
- Stodolsky, D. S. [1995]. Consensus journals: Invitational journals based upon peer review, *The Information Society* **11**(4): 247–260.
- Stone, M. [1961]. The opinion pool, *The Annals of Mathematical Statistics* **32**(4): pp. 1339–1342.
- Stvilia, B., Twidale, M. B., Smith, L. C. and Gasser, L. [2008]. Information quality work organization in wikipedia, *J. Am. Soc. Inf. Sci. Tech.* **59**(6): 983–1001.
- Suh, B., Chi, E., Kittur, A. and Pendleton, B. A. [2008]. Lifting the veil: improving accountability and social transparency in wikipedia with wikidashboard, *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI '08)*.
- Suh, B., Convertino, G., Chi, E. H. and Pirolli, P. [2009]. The singularity is not near: slowing growth of Wikipedia, *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym '09)*, ACM, New York, NY, USA, pp. 1–10.
- Sumi, R., Yasseri, T., Rung, A., Kornai, A. and Kertész, J. [2011]. Edit wars in wikipedia, *ArXiv preprint* .
- Surowiecki, J. [2004]. *The Wisdom of the Crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*, Doubleday.

- Tajfel, H. [1982]. Social psychology of intergroup relations, *Annual Review of Psychology* **33**(1): 1–39.
- Tapscott, D. and Williams, A. D. [2006]. *Wikinomics: how mass collaboration changes everything*, Portfolio.
- Taraborelli, D. and Ciampaglia, G. [2010]. Beyond notability. Collective deliberation on content inclusion in wikipedia, *Self-Adaptive and Self-Organizing Systems Workshop (SASOW), 2010 Fourth IEEE International Conference on*, pp. 122–125.
- Travieso, G. and da Fontoura Costa, L. [2006]. Spread of opinions and proportional voting, *Phys. Rev. E* **74**: 036112.
- Turner, J. C. [1989]. *Rediscovering the social group: A self-categorization theory*, Blackwell Publishers, London.
- Van Alstyne, M. and Brynjolfsson, E. [2005]. Global village or Cyber-Balkans? Modeling and measuring the integration of electronic communities, *Management Science* **51**(6): 851–868.
- Vancheri, A., Giordano, P., Andrey, D. and Albeverio, S. [2008]. Urban growth processes joining cellular automata and multiagent systems. part 1: theory and models, *Environment and Planning B: Planning and Design* **35**(4): 723–739.
- Vespignani, A. [2009]. Predicting the behavior of techno-social systems, *Science* **325**(5939): 425–428.
- Visser, B. and Swank, O. H. [2007]. On committees of experts, *Quarterly Journal of Economics* **122**(1): 372.
- von Ahn, L. and Dabbish, L. [2008]. Designing games with a purpose, *Commun. ACM* **51**: 58–67.
- von Ahn, L., Hopper, N. and Langford, J. [2005]. Covert two-party computation, *Symposium on the Theory of Computing (STOC)*, pp. 513–522.
- Von Neumann, J. and Morgenstern, O. [1944]. *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, USA.
- Voss, J. [2005]. Measuring wikipedia, *Proceedings International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden.

- Vázquez, A., Oliveira, J. G., Dezső, Z., Goh, K.-I., Kondor, I. and Barabási, A.-L. [2006]. Modeling bursts and heavy tails in human dynamics, *Phys. Rev. E* **73**(3): 036127.
- Viegas, F. B., Wattenberg, M. and Dave, K. [2004]. Studying cooperation and conflict between authors with history flow visualizations, *Proceedings of the SIGCHI conference on Human factors in computing systems*, Vienna, Austria, pp. 575–582.
- Viegas, F. B., Wattenberg, M., Kriss, J. and van Ham, F. [2007]. Talk before you type: Coordination in wikipedia, *Proceedings of the 40th Hawaii International Conference on System*.
- Viegas, F. B., Wattenberg, M. and Mckee, M. [2004]. The hidden order of wikipedia, *Proceedings of the 2nd international conference on Online communities and social computing (OCSC '07)*, Lecture notes in computer science, Springer, Berlin / Heidelberg, pp. 445–454.
- Wainwright, J. [2003]. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, John Wiley & Sons, Ltd., chapter Modeling and Understanding Human Behavior on the Web, pp. 170–209.
- Wang, F.-Y., Carley, K. M., Zeng, D. and Mao, W. [2007]. Social computing: From social informatics to social intelligence, *Intelligent Systems, IEEE* **22**(2): 79 – 83.
- Wash, R. and Rader, E. [2007]. Public bookmarks and private benefits: An analysis of incentives in social computing, *Proceedings of the American Society for Information Science and Technology* **44**(1): 1–13.
- Watts, D. J. [2004]. The “new” science of networks, *Annual Review of Sociology* **30**(1): 243–270.
- Watts, D. J. [2007]. A twenty-first century science, *Nature* **445**: 489.
- Watts, D. J. and Strogatz, S. H. [1998]. Collective dynamics of ‘small-world’ networks, *Nature* **393**(6684): 440–442.
- Weidlich, W. [1971]. The statistical description of polarization phenomena in society, *British Journal of Mathematical and Statistical Psychology* **24**: 251.
- Weidlich, W. [2000]. *Sociodynamics—a Systematic Approach to Mathematical Modelling in the Social Sciences*, Harwood Academic Publishers, Amsterdam, NL.

- Weisbuch, G., Deffuant, G., Amblard, F. and Nadal, J. P. [2003]. *Heterogenous agents, interactions, and economic performance*, Vol. 521 of *Lecture notes in economics and mathematical systems*, Springer, Berlin / Heidelberg, chapter Interacting agents and continuous opinions dynamics, pp. 225–242.
- Wells, H. G. [1938]. *World Brain*, London: Meuthuen & Co., Ltd.; Garden City, NY: Doubleday, Doran & Co., Inc.
- Wenger, E. C. and Snyder, W. M. [2000]. Communities of practice: The organizational frontier, *Harvard Businnes Review* **January-February**: 139–145.
- Wilkinson, D. M. [2008]. Strong regularities in online peer production, *Proceedings of the 9th ACM conference on Electronic commerce*, Chicago, Illinois USA.
- Wilkinson, D. M. and Huberman, B. A. [2007]. Cooperation and quality in wikipedia, *Proceedings of the 3rd International Symposium on Wikis and Open Collaboration (Wikisym '07)*, Montréal, Québec, Canada.
- Witte, E. and Davis, J. [1996]. *Understanding Group Behavior: Consensual Action by Small Groups*, Lawrence Erlbaum, Mahwah, NJ.
- Wood, S. N. [2010]. Statistical inference for noisy nonlinear ecological dynamic systems, *Nature* **466**(7310): 1102–U113.
- Wu, F. and Huberman, B. A. [2004]. Social structure and opinion formation, *ArXiv Condensed Matter e-prints* .
- Wu, F. and Huberman, B. A. [2007]. Novelty and collective attention, *Proceedings of the National Academy of Sciences* **104**(45): 17599–17601.
- Wu, F. and Huberman, B. A. [2008]. Public discourse in the web does not exhibit group polarization. Social Computing Lab, HP Labs.
- Wu, F., Huberman, B. A., Adamic, L. A. and Tyler, J. R. [2004]. Information flow in social groups, *Physica A: Statistical Mechanics and its Applications* **337**(1-2): 327–335.
- Wöhner, T. and Peters, R. [2009]. Assessing the quality of wikipedia articles with lifecycle based metrics, *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (Wikisym '09)*, pp. 1–10.
- Yang, J., Wei, X., Ackerman, M. and Adamic, L. [2010]. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities, *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Zachte, E. [2011]. Wikimedia report card – April 2011.

URL: http://stats.wikimedia.org/reportcard/RC_2011_04_detailed.html

Zhou, S. and Mondragón, R. J. [2004]. Accurately modeling the internet topology, *Phys. Rev. E* **70**: 066108.

Zickuhr, K. and Rainie, L. [2011]. Wikipedia, past and present, *Technical report*, Pew Internet & American Life Project. Last accessed: September 09, 2011.

URL: <http://www.pewinternet.org/Reports/2011/Wikipedia.aspx>

Zlatić, V., Božičević, M., Štefančić, H. and Domazet, M. [2006]. Wikipedias: Collaborative web-based encyclopedias as complex networks, *Phys. Rev. E* **74**(1): 016115.

Index

- activity lifespan, 9, 11, 43, 46, 65
- activity theory, 36
- Adam Smith, 24
- Adolphe Quet  let, 24
- AIC, *see* Akaike information criterion
- Akaike information criterion, 54
- Anderson-Darling, 51
- anonymity, 37
- assortativity, 26
- attitude change, 38
- Auguste Comte, 24
- average man, 24

- bond theory, 37
- bots, *see* Web robots
- bounded confidence, 16, 28, 39, 68
- Brooks' law, 20
- BSD, 5
- Burning Man, 35
- burstiness, 69, 70

- Canton Ticino, 38
- censoring, 47
- Ceresio, *see* Lugano
- Clement VII, 46
- Clickworkers, 23
- club theory, 40
- collaborative filtering, 16, 21–22
- common, 1
- common identity theory, 37
- community of practice, 41
- computerization, 18
- confidence, 68
- consensus, 27

- Conway's law, 20
- critical mass theory, 37
- crowdsourcing, 17

- Digg, 23

- edit cascade, 69
- edit filter, 3
- EM, *see* expectation maximization algorithm
- Encyclop  dia Britannica, 5
- ER, *see* networks, Erd  s-R  nyi
- expectation-maximization algorithm, 49

- Foldit, 22
- FOSS, *see* free open source software
- Free Open Source Software, 3, 5–7
- free open source software, 33, 35, 36
- free-riding, 35, 36

- Game Theory, 24
- games with a purpose, 23
- General Public License, 5
- gift culture, 35
- GNU, 5
- Gnutella, 36
- GPL, *see* General Public License
- graph, 25
- groupthink, 39

- H.G. Wells, 4
- HIT, *see* human intelligence task
- human dynamics, 17
- human intelligence task, 22

- inactivity, 44, 45, 47, 69
- incentive structure, 17, 34
- inter-edit time, 44, 45, 58
- Internet Relay Chat, 6, 37
- IRC, *see* Internet Relay Chat
- Ising's model, 24
- James Clerk Maxwell, 24
- Jimmy Wales, 3
- K-S test, *see* Kolmogorov-Smirnov
- Kolmogorov-Smirnov test, 51
- latent variable, 49–50
- likelihood, 55
- Linus Torvalds, 6
- locale, 45, 46
- log-likelihood, 49
- Lugano, 38
- lurking, 9, 44
- Mechanical Turk, 22
- Mediawiki, 3, 43, 46, 48
- meta-contrast, 38, 39
- meta-data, 9
- mixture model, 49–50
- Moore's Law, 19
- Nature, 5, 8, 42
- networks, 16, 25
 - epidemics, 27
 - Erdős-Rényi, 25
 - motifs, 26
 - Poisson random graph, 25
 - preferential attachment, 25–26, 70
 - small worlds, 25
- non-rival good, 35
- norms, 9, 66
- NPOV, *see* Wikipedia, neutral point of view
- Nupedia, 3
- open source, *see* free open source software
- opinion dynamics, 16, 39
- opinion space, 68
- over-confidence, 39
- P2P, *see* peer-to-peer
- PA, *see* networks, preferential attachment
- peer production, 1, 9, 15, 33, 36, 37, 39–41
- peer-review, 34, 36
- peer-to-peer, 36
- Poisson process, 69
- reciprocity, 34, 35
- recommendation, 21
- relevance filtering, 21
- remuneration, 34
- Requests for Adminship, 66
- SA, *see* sensitivity analysis
- sampling bias, 47
- secure sockets layer, 19
- selection bias, 47
- self-actualization, 36
- self-categorization, 38
- sensitivity analysis, 77
- social computing, 21–23
- social dilemma, 35–36
- social influence, 21, 39
- social informatics, 15–19
- social information processing, 39
- social judgment, 38
- social psychology, 37, 38
- socialization theory, 37
- socio-dynamics, 24
- socio-physics, 16
- socio-technical system, 16, 18–19
- software engineering, 20
- speed, 68
- SSL, *see* secure sockets layer
- STS, *see* socio-technical system

- task allocation problem, 40
- TGMM, *see* truncated Gaussian mixture model
- Thomas Hobbes, 24
- tragedy of the commons, 35
- truncated Gaussian mixture model, 48
- truncation, 47, 48, 50

- uncertainty, *see* speed
- universality, 11
- use case, 18
- user participation, 9, 40, 43

- Voter model, 24

- Web, *see* World-wide Web
- Web 2.0, 16, 21–23
- Web robots, 48, 51
- wiki, 3, 9, 35, 37
- Wikimedia Foundation, 43
- Wikipedia, 2–5, 17, 23, 36, 37, 39, 40, 42–45, 70
 - “revision” table, 46
 - accuracy, 40
 - anonymous editors, 44
 - civility, 66
 - consensus, 4, 41
 - data dumps, 46
 - deletion of pages, 67
 - good faith, 4, 66
 - governance, 40, 41
 - growth, 37
 - neutral point of view, 3, 4, 9, 65
 - rollback, 68
 - vandalism, 68
- wisdom of the crowds, 21, 39
- World Encyclopedia, 4
- World-wide Web, 16, 19, 32
- www, *see* World-wide Web

- Zürich, 38