Social Networks of Wikipedia

Paolo Massa
Bruno Kessler Foundation
Via Sommarive, 14
38123 Povo (TN), Italy
massa@fbk.eu

ABSTRACT

Wikipedia, the free online encyclopedia anyone can edit, is a live social experiment: millions of individuals volunteer their knowledge and time to collective create it. It is hence interesting trying to understand how they do it. While most of the scholar attention focused on article pages, a less investigated share of activities happen on user talk pages, Wikipedia pages where a message can be left for the specific user. This public conversations can be studied from a Social Network Analysis perspective in order to highlight the structure of the "talk" network. In this paper we focus on this preliminary extraction step by proposing different algorithms. We then empirically validate the differences in the networks they generate on the Venetian Wikipedia with the real network of conversations extracted manually by coding every message left on all user talk pages. The comparisons show that both the algorithms and the manual process contain inaccuracies that are intrinsic in the freedom and unpredictability of Wikipedia syntax and practices. Nevertheless, a precise description of the involved issues allows to make informed decisions and to base empirical findings on reproducible evidence. Our goal is to lay the foundation for a solid computational sociology of wikis. For this reason we release the scripts encoding our algorithms as open source and also some datasets extracted out of Wikipedia conversations, in order to let other researchers replicate and improve our initial effort.

Categories and Subject Descriptors

I.7 [**Computing Methodologies**]: Document and text processing

General Terms

Algorithms, Human Factors, Measurement

Keywords

Wikipedia, wiki, social network, empirical analysis, open source.

1. INTRODUCTION

Wikipedia, the free online encyclopedia anyone can edit, is a live experiment in collaboration and coordination among humans at large scale. The unpaid and volunteer work of millions of people was able to produce and maintain a public resource whose quality is comparable with more traditionally built encyclopedias [1]. On January 2011, Wikipedia is among the ten most visited sites of the entire web. The English Wikipedia started in 2001 and hence have been operational for more than 10 years. It is the largest Wikipedia with more than 3.5 millions article pages and globally received more than 440 millions edits. The registered users performing such edits are almost 14 million [2]. But the English Wikipedia is just the tip of the iceberg, in fact there are Wikipedias in 279 languages ranging from well developed communities such as the German and the French ones (with more than one million pages) to the Wikipedia in Latin or in Italian regional dialects such as Venetian. But wikis are not limited to online versions of encyclopedias in different languages. The Wikimedia Foundation runs also other multilingual projects such as Wiktionary (dictionary), Wikibooks (collectively written books), Wikinews (news). Moreover the open source software powering these projects, Mediawiki, is also used in many other commercial wikis, some of them with more than one million pages as well [2].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HT'11, June 6–9, 2011, Eindhoven, The Netherlands.
Copyright 2011 ACM 978-1-4503-0256-2/11/06...$10.00.

These very large collective products are produced by millions of people. However most of the academic efforts focused on the product itself, analyzing the quality of the content with human language technologies or, for instance, its category structure from a semantic web perspective, and somehow neglected the social side of these wikis.

We believe these online settings and communities offer a ripe area of research on interactions among humans. In particular, in this work, we focus on user talk pages. While Mediawiki software serves by default article pages, it also automatically assigns to every registered user a user page and a user talk page (henceforth UTP). For example, the English Wikipedia user "Mary" has a user page at http://en.wikipedia.org/wiki/User:Mary and UTP at http://en.wikipedia.org/wiki/User_talk:Mary. These are pages anyone can edit similar to article ones but have a special purpose. In particular, UTPs are intended for direct communication, meaning that any user can edit the UTP and leave there a personal message for its owner. Usually users end the message with their signature so that the receiver can check who wrote it. Figure 1 shows a UTP with messages and signatures.

These public conversations can be studied from a Social Network Analysis perspective: it is in fact possible to extract the network of whom "talked" to whom in the specific wiki simply by reading the messages. Social Network Analysis (SNA) is a well known technique in sociology. It is used for studying the structure of networks among individuals focusing both on the nodes themselves and on the relationships among them. For instance, a key concept in SNA is the centrality of the individuals which can map their relative importance within the network.

The large adoption of social networking sites and Web2.0 online communities such as Facebook has generated additional interest in SNA but also new challenges. While up to few years ago, networks under analysis were typically constituted by hundreds of nodes at most, now it is possible to measure and study networks populated by millions of users. For instance, Mislove et al. analyzed networks extracted from Flickr, YouTube, LiveJournal, and Orkut and their dataset contains over 11.3 million users and 328 million links [3].

Beside the very large dimension of networks, another even bigger challenge is the fact that the collection procedure and its assumptions highly influence the collected network and hence the findings that can be inferred from it. For instance, the previously mentioned paper [3] reports that the average number of

Figure 1: User talk page of user Phauly on English Wikipedia (http://en.wikipedia.org/wiki/User_talk:Phauly)

connections on Youtube is 4.29 while another published work [4] reports it as 6.80. The point here is that depending on how the network is collected, it is possible to infer very different results.

In this paper we propose algorithms for the extraction of social networks from conversations happening on user talk pages of Wikipedia and other wikis. More importantly, we empirically show that small changes in the collection procedures can produce diverse networks and hence different results, as in the case of average number of Youtube contacts. This is done by comparing the networks extracted from a small Wikipedia, the one in Venetian language, with the real network built by manually annotating each message left on this Wikipedia. A detailed description of the issues, their relative magnitude and ways to cope with them is provided.

The aim of this work is to lay down the basis for a solid computational sociology of wikis. For this reason, we release as open source the scripts encoding the proposed extraction algorithms so that other researchers can reuse them. We also release datasets extracted with the different algorithms.

2. SOCIAL NETWORK FROM USER TALK PAGES DISCUSSIONS

The focus of this paper are networks representing public conversations happening on user talk pages in Wikipedia. While user pages are written mainly or only by the owner of the page in order to briefly describe herself, UTPs are edited mainly by other users with the intention of leaving a public message to its owner.

Notwithstanding Wikipedia guidelines state that the site is not a social network nor a blog, we believe patterns of communication can shed an interesting light on the structure of the Wikipedia community and its internal functioning.

An example of user talk page is provided in Figure 1. It is important to note that a message is usually ended with the signature of the user who wrote it. While signatures are forbidden on article pages, users are encouraged to leave them at the end of messages they write on talk pages. There is a talk page associated to each article page and its aim is to discuss improvements to the article itself in order to reach consensus before making a change, instead of, for example, directly editing it. Leaving a personal signature is advised on talk pages and of course also on UTPs.

Wikipedia guidelines specify to add four tildes (~~~~) in order to sign a message. This can be done by typing directly the four tildes or by clicking on a signature button in the edit toolbar that appears when editing any Wikipedia page. The four tildes are automatically transformed by the software into a signature. A first important point is hence that identifiers of the writing user are not automatically inserted, as it happens for posts on Facebook walls for instance, but must be inserted by the user manually.

The default format for a signature, on the English Wikipedia, comprehends the username (as link to her user page) and the word "talk" (as link to her user talk page), in general followed by the date when the signature was left. Wikipedia in other languages have very similar formats for default signatures.

However users can personalize their signature via their site settings. Wikipedia guidelines state that the signature should not be distracting or confusing about colors and text and should contain no images. Moreover signatures must include at least one internal link to the user page, user talk page, or contributions page. For instance, the second message of Figure 1, starting with the headline "Wikipedia endnote assistant" was written by user Martin whose Wikipedia nickname is Smith609. This user personalized his signature in order to display both his real name and his nickname.

It is also possible that users forget to sign or they don't know they should do it or how to do it and this might result in missing or not correctly formatted signatures. Moreover links to user pages can be wrongly interpreted as signatures even when they are not. We will report on these specific issued in the empirical analysis described in Section 4.

When someone leaves a message for user A (i.e. edits the user talk page of user A), the interface show a "You have new messages" box the next time she signs in. This is the most visible way of knowing about new received messages.

The third message in Figure 1 is written by an anonymous user. In fact, users in Wikipedia have different roles. Users who don't authenticate themselves with a personal login and password are considered anonymous. If they contribute to the Wikipedia's content, they are identified by IP address of the computer from which they are connected. On the other hand, registered users can login in Wikipedia and their contributions are identified by their nicknames. Registered users can be elected to specific roles with additional powers such as system operators or bureaucrats. Bots are automated scripts created to carry out repetitive and trivial tasks such as checking for copyright violations, fixing small typos, reporting possible vandalisms and others.

The goal of this work is to create social networks that represent the conversation patterns among users of wikis (Wikipedia and other Mediawiki-based wikis) in order to then study and measure

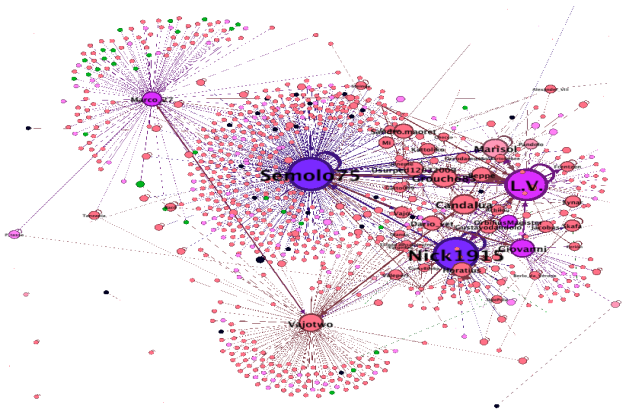

Figure 2: Social network of Venetian Wikipedia users based on UTP conversations. Node size is based on indegree.

them using Social Network Analysis. For example, processing the UTP in Figure 1 should discover 3 messages written to user Phauly from registered users Shell, Smith609 and the anonymous user identified just by its Internet address (217.77.80.29).

By processing all the UTPs in a certain wiki, it is possible to extract the complete conversation network. Figure 2 shows the network based on all the messages left on UTPs in the Venetian Wikipedia. SNA models networks as graphs in which nodes are typically people, and in our case wiki users, and edges are relationships. Here relationships represent conversation contacts. The network we extract from wikis is directed and weighted [3]: weighted because it is recorded not only the presence or absence of an edge but also its importance, i.e. each edge carries the number of written messages, directed because the edge from node A to node B, that represents how many messages have been written by A on the UTP of B, is kept separated from the opposite edge from B to A. Directed and weighted networks are richer of information than undirected and unweighted ones, for instance they can be studied from the perspective of the receiver (indegree as number of received messages) or from the dual perspective of the writer (outdegree as number of written messages). Figure 2 is the talk network extracted from the Venetian Wikipedia. It comprehends all the messages left from its inception in 2005 up to 2009, i.e. it is the cumulative network of messages written during 5 years. It is possible to analyze subsets of the network, for example by examining all the messages written in a certain year [5] or even to study them from a longitudinal point of view focusing on the evolution of the community [6].

Nodes of the network are Wikipedia users but only those who have been in some way active in user talk pages. Precisely, the network includes only users who left a message (present in at least the UTP of a user, in SNA terms with outdegree>1) and users who have their UTP created which could be because they received at least a message (indegree>1) but also in few cases because the users themselves created the page typically writing an information message specifying the user should be reached through another communication channel. In fact UTPs are not created by default.

2.1 Related work

Wikipedia is a large live social experiment whose electronic trails are made available by the Wikimedia foundation. As a consequence, recently many researchers analyzed Wikipedia in order to empirically test their hypotheses.

As already hinted, most of the attention concentrated on the actual content of the online encyclopedia by focusing on article pages while less attention have been devoted to users and their interactions. While most of the technologies employed to study article pages were linguistics, some researchers also used network analysis. For example, Capocci et al. [7] constructed a network in which nodes are Wikipedia articles and edges are links between articles. They analyzed topological properties of this network and proved the network growth can be described using the preferential attachment mechanism. Other researchers studied networks of articles [8, 9] typically to compute the relative importance of articles based on their centrality. Other papers focused on networks of Wikipedia categories instead of pages [10, 11].

The first paper to analyze messages left on Wikipedia, instead of article pages, is by Viegas et al, in 2007 [12]. They studied messages left on article talk pages by reading all message left in a sample of 25 talk pages and manually classifying them in 11 categories such as requests for information, or references to vandalism. Authors did not examine the pages where talks get archived when the talk page becomes too long. They describe the algorithms the coders followed to identify messages on talk pages and admit that talk pages are not easy to interpret given their free structure. For instance, they found that on average, users signed their names only 67% of the time. In addition to unreliable signature patterns, talk pages contain an enormous variety of postings which make it very hard to automatically and robustly parse them.

The focus in the work we present here is instead on user talk pages: UTPs (see Figure 1) can be compared to public emails while messages left on article talk pages are more similar to posts on forum topics. In fact, indentation was a key part of the algorithm used in [12] while, as we will see, on UTPs threads of indented conversations are less frequent. They did not construct networks of editing out of the talk pages and they would anyhow be about each single talk page.

In fact, Wikipedia user talk pages have been empirically analyzed even more rarely than article Talk page and not from a social network point of view. In [13] the first edit user A did on user B's talk page was considered as indicator of their first meeting. This point in time was used to compare the similarity of article pages the users edit and it was found that there is a sharp increase in the similarity between two editors just before they first interact, with a continuing but slower increase that persists long after this first interaction. "Governance in Social Media: A case study of the Wikipedia promotion process" [14] analyzes how each Wikipedia user voted for or against the election of other users. They also considered the amount of messages exchanged by two users on their UTPs and found a positive correlation between this amount and the probability of a positive vote. In “Beyond Wikipedia: Coordination and Conflict in Online Production Groups” [15], authors examined 6,811 different wikis investigating also communication on user talk pages reporting for instance the percentage of edits falling in this namespace across wikis.

The work closer to ours is [16] whose goal is to identify social roles in Wikipedia. Authors used both structural signature methods by examining the distribution of edits across types of pages and the structure of relationships between editors. However, differently from us, they considered ego networks, i.e. graphs centered around each specific user, in order to find patterns representative of the different roles.

There have also been other networks in which nodes were users, extracted not from user talk pages but from other actions performed on Wikipedia, in general from edits to article pages. For example, in [17], the maximal connected component of the undirected graph obtained by linking two users if they both edited at least six pages of the Japanese Wikipedia was extracted and analyzed as a directed graph by showing that the network exhibits a power-law degree distribution. Similarly, in “Analyzing the Creative Editing Behavior of Wikipedia Editors Through Dynamic Social Network Analysis” [18] an edges from user A to user B was added to the network if user A edited a certain article page after B and different editing patterns were used to identify mediators, zealots, coolfarmers and egoboosters.

Other authors focused on reverts. Wikipedia offers a functionality by which a user can rollback a contribution by another user, reverting the text of the article page as it was before the contribution. Such revert graph [21, 22] is another example of network extracted from Wikipedia in which nodes are users but the edges represent a sort of negative relationship, for instance nodes that are more central are the ones whose contributions are less appreciated. In [19], they also note that “user talk pages were very useful in understanding disagreements and conflicts between users. We believe further effort in modeling user talk pages would provide a deeper insight into the social dynamics of Wikipedia”.

Another social network extracted from Wikipedia is based on who voted for or against whom in elections to become administrator of Wikipedia [14]. This network is smaller since few users participate in elections with their votes and edits.

The work presented here is different because it describes a methodology for extracting social networks out of conversations happening on user talk pages of Wikipedia and wikis in general. More importantly we discuss in details how different algorithms can produce different networks and compare the resulting networks. Since the goal is to start a solid global effort towards a computational sociology of wikis, we also release the scripts and the networks extracted for comparative purposes and future improvements by any interested researcher.

In this section we described what are user talk pages and how they are used, and our interpretation of UTPs for the construction of social networks representing communicative patterns among wiki users. We also reported related works pointing out how our work is different. In the next section we describe the methodology and algorithms we propose for extracting social networks from Wikipedia.

3. DATA COLLECTION AND EXTRACTION ALGORITHMS

Wikipedia has a very open nature: every page can be edited by anyone, even anonymously. The license of every Wikipedia page (including user talk pages) is Creative Commons Attribution-ShareAlike, that is very permissive since it basically allows to copy, distribute and modify each page as long as attribution is given and the license for the modified work is not changed. This feature is very important since it gives researchers the possibility to study the entire history of Wikipedia actions and also to release the collected datasets so that other researchers can verify empirical findings and improve the analyses.

Wikipedia is powered by Mediawiki and the open source nature of this software is one of the reasons behind the fact many other wikis use the same software for serving their communities. For instance, on Wikia.com it is possible to create for free a wiki about any topic and it is powered by Mediawiki. As January 2011, there are more than 50,000 wikis created on Wikia; 14 of them are very active having received more than one million edits and range from wikis for collecting song lyrics to wikis about games such as World of Warcraft, from wikis to collect answers to every question to wikis entirely based on nonsense (uncyclopedia) [2]. There are also hundreds of independently installed wikis powered by Mediawiki [2].

Mediawiki software includes a functionality for creating different XML files reporting every action ever performed on the system by every user. The Wikimedia foundation, which manages Wikipedias and other related projects such as Wikiktionary or Wikibooks, provides almost weekly the current XML files for everyone to download and analyze at http://dumps.wikimedia.org. Also commercial entities such as Wikia.com follow the same approach and these means that the number of available dumps for different communities count at least in the thousands [2].

The algorithms and methodology we present in this paper work for each of the wikis powered by Mediawiki that release the XML files about the activity occurred in the wiki. This fact allows to study small wikis just as easily as the largest ones, opening the possibility for a comparive study of communities developed on different wikis, communities with different goals, policies and rules, quantity and types of users, years and levels of activity.

Note that we have decided to treat every wiki separately, in fact, while a recent feature of Wikipedia gives the possibility to unify usernames across different Wikipedias, at the moment it is not easy to reliably identify if, for example, user “Mary” on English Wikipedia is controlled by the same person that controls user “Mary” on Italian Wikipedia. We comment on this comparison of behaviour of the same user across wikis along with possibility of studying the evolution of the networks at different timestamps in section 5. For instance, the English Wikipedia started in 2001 and hence it offers the opportunity to study the activities over 10 years of millions of users, almost 14 millions on January 2011.

In the following we describe how to extract the cumulative network of all messages exchanged on UTPs in one specific Wikipedia. In order to parse the XML files and extract the talk network, we developed a collection of scripts written in the programming language Python. Scripts are licensed as open source and are available at http://sonetlab.fbk.eu/data/. As we describe later, these XML files can be very large and this makes processing them automatically a non trivial task; for example a complete dump of just the English Wikipedia is an XML file of 5,600 Gigabytes of data. Sharing our scripts can help other researchers to verify and improve our research without the need to worry about coming up with ways to process these very large files. At the same web location we also share some datasets extracted with the scripts so that researchers can evaluate hypotheses on already available datasets and not to worry about processing them, which might require computational experience and lots of computational power and time.

3.1 Manual social network extraction

In the next sections we present the algorithms we propose for extracting social neworks out of conversations happening on user talk pages. However, in order to evaluate the performances of the algorithms and the issues in automatically parsing XML files of wiki activity, we decided to build a controlled network acting as ground truth that can be compared with the other extracted networks. For this reason, two colleagues analyzed by hand every message left on every user talk page of the Venetian Wikipedia.

We chose the Venetian Wikipedia because it is relatively small since, at January 2010, it counts 8,838 article pages and 6,634 registered users, compared for instance to the English Wikipedia with 3,548,48 article pages and 13,894,042 users. Our goal in fact was to manually extract the complete and real network and the relatively small size of the Venetian Wikipedia allowed us to do it.

First we queried Wikipedia server in order to get all the pages in the user talk namespace for the Venetian Wikipedia. Let us note again that the user talk page for a registered user is created only if someone edits it, in general in order to leave a message. User talk pages can exist also for anonymous users since it possible also to leave messages to anonymous users.

The two coders visited each page in the user talk namespace, both for registered and anonymous users. They divided each UTP into messages, according to the presence of a signature in the text, to the level of indentation of the text or even to the presence of a header with an horizontal row above the text, similarly to the procedure used in [12]. UTPs were analyzed as they were at December 30, 2009, since we are going to compare them with networks extracted from XML files created at that time as well.

For each identified message, coders recorded the owner of the UTP, the receiver, and the writer of the message as identified by the user page linked in the signature, the sender. In terms of Social Network Analysis, for each message written by user A on user B talk page, the two users were added as nodes to the network and the corresponding edge from A to B was created, if not already present, otherwise the weight on the edge,

```
<page>
    <title>User talk:Phauly</title>
      <revision>
        <text xml:space="preserve">
== '''Welcome!''' ==
Hello, {{BASEPAGENAME}}, and [[Wikipedia:Welcome,
newcomers|welcome]] to Wikipedia! Thank you for your
contributions. I hope you like the place and decide to
stay. Here are some pages that you might find helpful:
*[[Wikipedia:Five pillars|The five pillars of
Wikipedia]]
*[[Wikipedia:How to edit a page|How to edit a page]]
*[[Help:Contents|Help pages]]
*[[Wikipedia:Tutorial|Tutorial]]
*[[Wikipedia:Article development|How to write a great
article]]
*[[Wikipedia:Manual of Style|Manual of Style]]
I hope you enjoy editing here and being a
[[Wikipedia:Wikipedians|Wikipedian]]! Please
[[Wikipedia:Sign your posts on talk pages|sign your
name]] on talk pages using four tildes
(<nowiki>~~~~</nowiki>); this will automatically produce
your name and the date. If you need help, check out
[[Wikipedia:Questions]], ask me on my talk page, or
place <code><nowiki>{{helpme}}</nowiki></code> on your
talk page and someone will show up shortly to answer
your questions. Again, welcome! .
[[User:Shell_Kinney|Shell]]
<sup>[[User_talk:Shell_Kinney|babelfish]]</sup> 15:29, 7
November 2006 (UTC)

== "Wikipedia endnote assisstant" ==
Hi, sorry to take so long to reply to your message. It's
convention at Wikipedia to leave new messages at the
bottom of the page, and as I was moving country at the
start of September, I didn't see your message until now!
Have you tried the updated URL,
http://toolserver.org/~verisimilus/Scholar ? Let me know
if you continue to encounter problems.
Glad you find the tool useful! Best wishes,
[[User:Smith609|
Martin]]&amp;nbsp;'''<small>([[User:Smith609|
Smith609]]&amp;nbsp;-&amp;nbsp;[[User_talk:Smith609|
Talk]])</small>''' 01:19, 7 October 2008 (UTC)

== Test anonymous edit ==
Just a test done by myself on signature formatting. --
[[Special:Contributions/217.77.80.29|217.77.80.29]]
([[User talk:217.77.80.29|talk]]) 12:08, 8 February 2010
(UTC)
        </text>
    </revision>
</page>
```

Figure 3: Fragment of pages-meta-current XML file (only relevant tags) for UTP in Figure 1.

representing the quantity of messages A wrote to B, was increased. As an example, by referring to the UTP in Figure 1, the coders would identify three messages left to user Phauly by users Shell_Kinney, Smith609 and an anonymous user. It is possible that a user leaves a message on her UTP for replying to a message left by another user. This procedure is not very frequent as we will comment later and would result in a self-edge.

All messages that have been placed in archive pages were also coded. When the user talk page becomes too long, Wikipedia guidelines suggest to archive pages in a subpage, such as http://vec.wikipedia.org/User_talk:Phauly/Archive3.

We have already commented on how users leave their signatures on wiki page and how these can be personalized. Moreover signatures have to be manually inserted so there are cases in which the writer forgets to sign her message. The procedure followed by the coders was to look for signatures, even when they were not correctly inserted in the user talk page and, if it was reliably possible, to associate it with the user who wrote it. They hence coded the presence or absence of signatures and the fact that signature is formatted according to Wikipedia guidelines so that an automatic script can reasonably detect them.

As already explained, we decided to limit each network to a single Wikipedia. This means that, for example even if a user signed in the Venetian Wikipedia with a link to her username on English Wikipedia, we considered this user as not identifiable.

The manual coding process was part of a larger work in which the coders also collected the intention and other characteristics of the messages, in order to understand how UTPs are used. The manual coding is of course not feasible on every message of larger wikis such as the English Wikipedia and in fact the reason behind the manual coding was also looking for patterns, regularities and deviations from them in order to devise automatic scripts able to replicate the performances of human coders.

Thus the manually created network can act as the ground truth used to measure the quality of algorithms for automatically extract networks that we introduce in the next sections.

3.2 Automatic social network extraction from signatures on user talk pages

In this subsection we present an algorithm that basically mimic the behaviour of coders while the next subsection introduces an algorithm with a different approach.

The "signatures" algorithm is designed to reproduce what human coders would do (counting signatures left on UTPs) but automatically and hence much faster and on larger wikis. The algorithm coded in the Python script we released as open source (named signature2graph.py) takes in input the XML file produced by Mediawiki called pages-meta-current. Figure 3 shows a fragment of this XML file representing the user talk page of User Phauly on English Wikipedia whose rendering was presented in Figure 1. This XML file contains the text of the current version of every page in Wikipedia. The signature script parses this file by discarding all the pages (XML tag <page> in Figure 3) but the ones whose title starts with "User talk:" or the equivalent formulation in each language which is available at the beginning of the XML file, for example in Venetian Wikipedia this is "Discussion utente:". For each of these UTPs, the text content is parsed looking for signatures as regular expressions.

As already reported, according to English Wikipedia's guidelines, signatures must include at least one internal link to user page, user talk page, or contributions page. For instance, the wiki syntax of user "A" default signature in English Wikipedia would be `[[User:A]] ([[User_talk:A|talk]])`, which is rendered as "A (talk)", where "A" is a link pointing to A's user page and "talk" a link to her UTP.

In Figure 3 it is possible to observe the wiki syntax of the different user signatures which appear in the rendering of the UTP in Figure 1. For example, `[[User:Shell_Kinney|Shell]] <sup>[[User_talk:Shell_Kinney|babelfish]]</sup>` is just a little bit personalized with respect respects to Wikipedia guidelines: Shell is the name that appears in place of the link to the user page and babelfish is the text for the link to the user talk page.

In fact it is common that users personalize their signatures and coding messages on the Venetian Wikipedia highlighted dozens of different patterns. As a consequence, we decided to look for a minimal indicator of signature: our script detects links to user pages since they seems to be the most regular and recurring indicator of signature. In the previous example, the regular expressions of our script look for the presence of the word User (or the equivalent in the specific Wikipedia language) preceded by "[[" and followed by "|" and would extract the following text, as Shell_Kinney. There are a lot of irregularities in signatures and our script is very robust, for example about the presence of spaces, embedded HTML tags, non balanced parentheses. For the specific details of how signatures are detected, we refer to the regular expression code of the script signature2graph.py. Signatures of anonymous users include a link to her contributions page, as it is possible to see in the last signature of Figure 3.

```
<page>
  <title>User talk:Phauly</title>
  <revision>
    <timestamp>2006-11-07T15:29:48Z</timestamp>
    <contributor>
      <username>Shell Kinney</username>
    </contributor>
  </revision>
  <revision>
    <timestamp>2008-10-07T01:19:54Z</timestamp>
    <contributor>
      <username>Smith609</username>
    </contributor>
  </revision>
  <revision>
    <timestamp>2010-02-08T12:08:19Z</timestamp>
    <contributor>
      <ip>217.77.80.29</ip>
    </contributor>
  </revision>
</page>
```

Figure 4: Fragment of stub-meta-history XML file (only relevant tags) for UTP in Figure 1.

3.3 Automatic social network extraction from edit history of user talk pages

The second algorithm we propose does not try to mimic human coders but follows a different approach. While the first two proposed extraction methods ran on the current version of Wikipedia, this algorithm takes into account the whole edit history since the creation of the considered Wikipedia and hence it is called "history" (the Python script is utpedits2graph.py). The input of this algorithm is in fact the XML file called stub-meta-history. This file contains the revisions made on all the pages in the considered Wikipedia along with the user who was responsible for the edit and the timestamp but does not contains the text of the page. Figure 4 shows a fragment of this file corresponding to the related part of "pages-meta-current" of Figure 3 and the rendered UTP page of Figure 1.

Similarly to the previous one, the "history" algorithm considers only pages in the user talk namespace. For each edit (tag "revision" in the XML file) to the UTP of user A, the author of the edit is extracted and this is considered as a message from this user to user A. In order to chronologically limit the study of the community, a parameter of the script allows to specify to consider only edits done before or after certain dates.

Since in order to leave a message to user A the common actual practice is to edit the UTP of A, this algorithm does not look on indicators of signatures in the current UTP pages but analyzes the history of the UTP itself and operationalizes edits to it as messages written to the owner of the UTP.

Clearly, given this difference in the operationalization, the extracted networks might be different. In fact, the next section presents the results of the extraction of social networks from UTPs of the Venetian Wikipedia done manually by the coders and automatically by the algorithms signature and history, discussing their different issues.

4. COMPARISON OF EXTRACTED NETWORKS AND ISSUES

In this section we analyze in details the differences among the social networks of Wikipedia as they are extracted by the different algorithms previously described. The goal of our work is to start a reliable and reproducible study of the community of users in wikis. For this reason, being able to stand on clearly defined algorithms is a strong prerequisite.

The first network we extracted was the one created by human coders from Venetian Wikipedia. In fact the manual coding provided the possibility to compare results of the scripts with the ground truth of a more reliable network extracted by hand by humans through their judgment. If we were to run our algorithms on the very large English Wikipedia for instance we would have had no easy way to check for errors or possible improvements of the algorithms. On the other hand, even if practices observed on the Venetian Wikipedia are not straight away generalizable on other wiki communities, coding by hand the pages gave us the opportunity to notice patterns and regularities just as exceptions to them. Our contribution goes in this direction providing empirical evidence of the reliability of the extracted networks.

All networks were extracted as they were at December 30, 2009: coders visited pages as they were at that date, the signature algorithm was run on an XML file created on that date and the history algorithm received a parameter indicating to consider edits up that date.

In the following we discuss the most important issues in detail, focusing on the impact they have on the extracted networks.

4.1 Number of nodes

The most basic information in a social network is the number of nodes: in our case, each node represents a specific user and the different algorithms should be able to identify the same number of nodes in the networks. However this is not the case.

Table 1 shows that coders extracted a total of 918 users while the signature algorithm found only 906 and the history algorithm a larger number, precisely 981. Let us recall here that these networks only contain users who have been somehow active in conversations, either because their UTP has been created by someone or because they left at least a message on some other user talk page.

The differences are not extreme but they are still not trivial and already show how three different ways of creating a network out of the same wiki community produce three different networks. There are different reasons for these differences and we describe them in the following, commenting on how much they present an issue and influence Social Network Analysis results and findings.

4.2 Renamed users

Table 1 shows that different algorithms find a different number of users. A small issue but with an impact with this regard is the fact users in Wikipedia can ask to change their usernames. Renames are carried out by bureaucrats (users with additional powers). In the 5 years of activity of Venetian Wikipedia, there were just 15 cases of renaming, while in English Wikipedia there were 17,096 rename user activities up to 28 July 2010 recorded on a log file.

We analyzed the specific cases in the Venetian Wikipedia in order to gather insights about how these important changes for SNA are recorded and made visible in the XML files.

There are essentially two cases: (1) a person arrives in a specific Wikipedia and reclaims the username already used by another person, and (2) a person wants to have its username renamed. The first case is the most difficult to manage because there are two persons owning in different times the same username. The second case is easier to handle because the old username history is moved to the new name and the old username is not used anymore.

On Venetian Wikipedia, for example, the first case happened when the person owning the username "Maximillion Pegasus" in other Wikimedia projects asked the bureaucrats of Venetian Wikipedia to have the same username also in this wiki. Since the person controlling this username in Venetian Wikipedia was not active, the rename was performed: bureaucrats renamed "Maximillion Pegasus" into username "Usurped12032009". The user page and the user talk page (and their histories) were moved to the new name: in this way at the UTP of user

Table 1: Number of nodes and of edges for networks extracted from Venetian Wikipedia using different algorithms. The last three lines refer to the network without anonymous users and Marco27Bot as writers of messages.

	#nodes	#edges
Vec coding	918	1073
Vec signature	906	1087
Vec history	981	1869
Vec coding (no anon, no bot)	902	1056
Vec signature (no anon, no bot)	891	1062
Vec history (no anon, no bot)	882	1077

Usurped12032009 it is possible to read the messages received by the user also when its username was "Maximillion Pegasus". In fact, the "old Maximillion Pegasus" used the talk feature intensively and both received many messages on its talk page and left many messages on other user talk pages before becoming inactive while the "new Maximillion Pegasus" never wrote or received a message in Venetian Wikipedia.

Unfortunately, existing signatures are not affected by a rename. This means that both human coders and the signature algorithm would find a lot of outgoing edges in the talk network from Maximillion Pegasus (messages left by the previous user) and zero incoming edges (since the current UTP is empty). On the other hand the user Usurped12032009 would have a large indegree (number of received messages, when the user was named differently) and zero as out degree (number of written messages, since it never signed a message as Usurped12032009). Detecting this issue was not easy because it required cross checking XML and log files.

The history algorithm is not affected by this issue since in the meta-history XML files, the user responsible for the edit is the correct one, for example the edits made by Usurped12032009 when its username was "Maximillion Pegasus" are listed as made by Usurped12032009.

We might be tempted to consider this issue marginal but in reality more than 17,000 renames happened in the English Wikipedia and usually involving very active and peculiar users, and this issue affects the most basic element of social networks, that is number of nodes. As we have explained, this issue does not affect the history algorithm while it is hardly detectable even by coders.

4.3 Number of edges

The number of edges is the other basic characteristics of a social network graph. In our case it corresponds to the number of ordered pairs of users among which occurred at least a directed communication, i.e. a message written on the user talk page.

Table 1 reports this quantity for the three different extracted networks. The one extracted by coders counts 1073 edges, while networks built by the signature algorithm has 1087 edges. The history algorithm, similarly as it was for number of nodes, find a larger number of edges, precisely 1869.

4.4 Information messages and redirects

Some persons are active only in one or two Wikipedias and don't check their UTPs in the other ones. For example, the already cited user Maximillion Pegasus wrote an information message on top of own UTP, signaling to leave messages on a different wiki. On Venetian Wikipedia, these information messages are sometimes left using a template such as `{{softredirect|:en:User talk:Synthebot}}`. Templates are used for inserting common pieces of text but can be created by anyone and hence can be different in different wikis. The signature algorithm detects the one we found on Venetian Wikipedia but might fail on other wikis. Hard redirects have a similar aim but, being in wiki syntax (such as `#REDIRECT [[User talk:Alice]]`), the Mediawiki software automatically redirects the visitor to the target page.

Hence redirects are a way to associate the usernames of the same person on different wikis and could open the way to a study of users across wikis: instead of considering each wiki separately, it would be possible to analyze the global network of all wikis. However we decided not to do so because the process is highly unreliable. We already commented about the feature of unified logins which is still not too formalized in Wikipedia. Moreover reliably parsing usernames across every wiki (from Chinese Wikipedia to a commercial wiki on Wikia.com) would introduce additional biases and errors so we decided to be conservative and rely on analysis of communities of users inside each specific wiki.

Coders were instructed to count redirect and information messages. Out of 1786 coded messages on Venetian Wikipedia, 60 were information messages and 27 were redirects. Note that some of the information messages are signed and the signature algorithm would detect them, even if they would be messages from user A to user A, i.e. self-edges as we describe in next section. Similarly the history algorithm detects an edit of user A on her user talk page.

4.5 Messages to oneself

Users sometimes edit their own UTPs. This happens for example when a user decides to reply to a message directly on her UTP instead of leaving it in the UTP of the desired receiver. We have already commented that this behaviour does not trigger the alert about new messages received and so it is more likely to go unnoticed. The other case in which a user edits her UTP is typically for inserting a redirect or information message as we have already described.

In terms of Social Network Analysis these activity would result in self-edges, directed edges from node A to node A. But clearly these are not messages that the user wrote to herself.

While it could have been often possible for human coders to understand to which user the message written by A on her own UTP was addressed, it would be very hard to devise an automatic algorithm able to do so, mainly because Wikipedia pages are totally free in structure, so even if often a reply is indented with respect to the original message, this practices is not followed by all the users.

Our goal being to automatize the extraction on larger Wikipedias such as the English one, we are interested in understanding how prominent is this pattern of replies. Only 56 of the messages coded manually were written by User A on his/her UTP as a reply. This communication mean was used mainly by users who wrote a single message and this was a reply to the initial welcome message. While it is not possible to generalize this finding on other Wikipedias, this relatively small percentage provide some evidence of the fact that self loops are not a very important issue since they are not very frequent.

In order to mitigate this issue, the history algorithm could use some heuristics. For example, if UTP of A is edited by A after have been edited by B, possibly this is a message from A to B so an edge from A to B could be added even if the UTP of B was not edited. However this heuristic would add an additional level of unreliability and we decided not to follow this way. In fact sometimes users edit their UTP for fixing typos and grammar

errors or to style it according to wiki syntax. We believe the most conservative choice is to remove self edges from the analysis, especially because their presence as real but hard to detect messages is minimal.

4.6 Non human users writing messages

Bot are non-human users, granted to perform automatic actions within Wikipedia. The name come from "robot" to point out their nature. Bots are widely used to fight spam and vandalism, correct small grammatical errors and make in a batch a whole set of changing (e.g. changing a template throughout all the pages). Especially on article pages, bots perform a significant part of the edits [23]. In Wikipedia, there is no visible difference between human users and bots, for example both of them have user and user talk pages. A commonly adopted convention is to include the word "bot" in the username of the bot.

With respect to redirects and information messages, it should be noted that often it is the human controller of the bot that writes them, for signaling that this user is not an human able to reply to messages and messages should be addressed to the controller. In this case the automated algorithms would detect that some user wrote a message to the bot and it could possibly be incorrectly inferred that there is social activity ongoing among humans and not humans on Wikipedia.

Another very important issue with bots is that they can be used to write almost automatic messages that are signed with non-bot user signatures. For example, when a new user firstly login, in many Wikimedia projects she receives a welcome message, usually linking to some of the most important project's guidelines. An example of welcome message is the first one in Figure 1. Policies about his are different in each wiki. For instance on Italia Wikipedia there is a bot, "Benvenuto bot" which automatically edits new users talk pages by leaving a welcome message signed with the signature of users who volunteered to be identified as the welcomers. Instead on English Wikipedia welcomes are given manually by users.

Out of 1786 messages extracted by coders, 774 (43.33%) were welcome messages, often written as templates. Both coding and signatures' algorithms assigned those messages to the user who signed them, for a large part by project's administrator. History algorithm however spotted the true author of the post, i.e. the bot. In the case of Venetian Wikipedia a bot, called Marco27Bot, became operational at the end of 2009 with the task of welcoming new users and this is responsible for the larger number of edges detected by the history algorithm.

The empirical evidence that we accumulated by analyzing by hand the Venetian Wikipedia is that hypotheses involving bots must be carried out being aware of the specific behaviour of each bot and the period in which it was operational, as for example in [23] which focused just on the consequences on users of the activity of one specific bot. With this regard, ad additional script we wrote and released performs the enrichment of the network by adding specific information such as the role of the user (node of the network) so that specific nodes such as bots or anonymous can be excluded from the analysis, if required.

4.7 Anonymous users, vandalism and deleted messages

Anonymous users are identified by their Internet address: they can leave messages with their signature and also have a user page and a user talk page so that they can also receive messages.

UTPs of anonymous users in the Venetian Wikipedia received mainly automatic messages, for example welcome messages. They also received most of the messages signaling a possible vandalism, that is 33. Only 8 warning messages were written to registered users. Both signature and history algorithm detect all the messages received by anonymous users because the signature and edit history are very regular.

It is more interesting to analyze the active behaviour of anonymous users, i.e. the message they wrote. Coders found only 9 messages written by anonymous users. However the history algorithm found many more anonymous users writing messages. In fact, even if Wikipedia guidelines suggest not to delete messages, sometimes users do it, especially when they receive vandalism. Vandalism messages tend to be written by anonymous users so this explain why many anonymous users cannot be detected by coding and signature algorithm. Besides anonymous users, also three registered users wrote only vandalism messages which were removed from UTPs so that they were present only in the network extracted by history algorithm. Excluding anonymous writers makes number of nodes in the three different networks more similar (see Table 1) and the extraction process more reliable.

4.8 Many edits per message

Networks created by coding and by signature algorithm are based on the current version of Wikipedia pages, whereas the history algorithm parse the edit history of pages. Hence, one important difference in the network generated by the two approaches is given by the fact that sometimes a user edits the same message more than once, for example if she discovers a small typo after the message is saved. In this case, the history algorithm would detect two messages, resulting in an incorrect over representation of the communication pattern.

This is one of the reasons for the fact history algorithm builds a network with a larger number of edges than the other procedures. A possible fix to this is to considered messages written by A to B in a short time window as a single message instead of many messages. However we did not include this functionality because it is not easy to find a time window value that would fit all wiki communities and the extracted networks will depend on the choice of this arbitrary value.

4.9 Personalized, missing or incorrectly formatted signatures

One of the most important characteristics of the signature algorithm and the manual coding is the process by which signatures are identified. As already reported it is possible to personalize signatures in Wikipedia and guidelines express how far from the default it is possible to personalize. By looking at signatures in Venetian Wikipedia we noted that there is a large variance in personalized signatures. So we decided that the most robust and reliable way to detect signatures was to look for links to user pages. Note that this heuristics, embedded in the signature algorithm, does introduce errors, for example the personalized signature of user Smith609 (see Figure 3 for the wiki syntax) contains two links to the user page and would be detected twice.

Personalization is an important issue especially for very active users. They typically change their signature conspicuously and of course, for the significance of the network, it is more problematic not to detect the signature of a very active user than of a more marginal user which might have received just one message. Moreover they also change their signature many times during the years so that their activity might be detected for a period and not detected for another period, leading to conclusions about users leaving and returning which might be non correct. For example one of the most active user in the Venetian Wikipedia, Nick1915,

very central in the network as Figure 2 shows, exhibited a small outdegree in the signatures network. This happened because for a certain period he signed messages using a template rather than the usual signature. The template is `{{Utente:Nick1915/firma}}`. At the moment templates are interpreted by our scripts since it is extremely slow to do this on large wikis, considering that there are thousands of templates and that templates can translude other templates generating long and possibly infinite transclusion chains. For the specific case of the Venetian Wikipedia, this template was included as valid signature for the signature algorithm but of course this is not robust since on other wikis the templates used for signatures could be different.

It is also possible that users forget to sign or they don't know they should do it or how to do it. Out of 1786 messages extracted by coders, 148 (8.3%) were signed by users who were not detectable. For the largest part, the reason was that the users didn't leave their signature under the message, so the messages were unsigned. The coders tried to associate the signatures to users of the Venetian Wikipedia when the signature was present but for example there was no link to the user page. However, as already reported we decided not to associate users who signed with hyperlinks to identities on Wikipedias other than the Venetian one. They were also 6 occurrences of users who signed simply with their names but no hyperlinks such as "beppe" or "Lucia da Zurigo" and these were not associated to Wikipedia users.

On the other hand, the history algorithm does not require a message to be signed in order to discover its author, since the editor is provided within the XML file. Moreover in case of renamed users, the signature is the correct one.

Note that on some Wikipedias (notably the English one but not on Venetian one), there is a bot, called Sinebot, which automatically inserts signatures when a user forgets to leave her signature after an edit. Interestingly this behaviour have sparked some discussions by people who protested their lack of signature was intentional, as reported in [23]. The presence of this bot, heavily affecting signatures, again suggests that bots should be treated separately with a specific knowledge about their automatic behaviour. We did not find examples of users signing with the username of another user, beside the already explained case of owners of bots. Hence we are not in position to comment on the fact this is an issue, for example on larger Wikipedias. Moreover the mere presence of a link to a user page would be detected by signature algorithm as a message even if it was simply a pointer to that page written inside a normal message. The human coders did not find any occurrence of this pattern.

4.10 Date of message

Signatures created with 4 tildes are rendered with the date of the signature at the end (see Figure 1). This opens the possibility to analyze networks from a longitudinal point of view by adding the date of the message as an attribute over the edge. Note that the signature algorithm would have to do it by parsing how dates are written in the specific language, for example in the Venetian Wikipedia the months are written in the language equivalent and this format can change over time. On the other hand, the history algorithm relies again on the precise format written in the XML file (tag timestamp on Figure 4) that is standard for all wikis.

4.11 Archived messages

When a user talk page become too long, guidelines suggest to archive it. Typically archive pages are created as subpages of the UTP such as `User_talk:Phauly/Archive3`. However anyone is free to create subpages for different purposes under the UTP. Usually archives are linked from the UTP and hence human coders were able to follow the links and to extract archived messages left on these pages as well.

On the other hand, the signature algorithm has to rely on heuristics for deciding if a certain subpage of UTP (appearing as a separate page in the XML pages-meta-current file) is an archive of messages or has another purpose. The heuristics encoded in the algorithm is to look for the presence in the page title of “vecchi”, “archiv” or “old” which were the ones human coders detected on Venetian Wikipedia. However this process is evidently not robust with respect to the language and hence not reproducible, for example in the Chinese Wikipedia messages might be archived in subpages with different titles.

The history algorithm is not affected by this issue since it consider only the main UTP and not its subpages. In fact the history of edits remains associated to the main UTP and hence there is no need to process subpages.

5. DISCUSSION

In the previous section we have described the main issues involved in the extraction of a social network of communication out of UTPs in wikis. Notwithstanding the availability of data which does not require to crawl web sites with heuristics and even on a community as small as the Venetian Wikipedia, different algorithms produce different networks. Moreover it is not easy to argue what is the correct network and it depends on the perspective under which the hypotheses are posed. Even basic questions such as number of nodes and number of edges can receive different answers.

Overall the differences between the network built analyzing by hand UTPs and the one extracted automatically by the signature algorithm are not very large (see Table 1). It could be argued that the second can't be better than the first but this is not true because manual coding is an error-prone and non-reproducible process.

The largest differences are with respect to the network produced by the history algorithm which adopts a different approach to UTPs not looking for signatures but parsing their edit history. In fact the number of nodes and edges seems much larger (see Table 1). However, as already commented in previous section, the history algorithm also detects edits made by anonymous users which are later deleted because they are considered vandalism. This is the reason for the difference in number of nodes.

Moreover edits made by bots are often not visible on UTPs but are detected by the history algorithm. In the specific case there is a bot called Marco27Bot that was not found by human coders and signature algorithm since it never left a signature but edited 684 UTPs of other user. In fact 373 of them were performed in one single day (September 6, 2009) in order to replace welcome templates left earlier by registered users. After this date, Marco27Bot performs the function of the already commented Benvenuto bot in the Italian Wikipedia of welcoming new user by appending the signature of two real registered users who volunteered to be identified as welcomers.

This point clearly highlights that the network extracted are different in intentions: one can be interpreted as the network users see (with its variability in signatures and formats), the other as the network of what really happened in the wiki. However, as Table 1 shows, by not considering anonymous users and bots as writers, the differences are reduced significantly. Studying the behaviour of anonymous users as writers of messages would deal with hypotheses about vandalism and would require specific heuristics in the extraction process. Similarly studies on behaviours of bots such as [23] require ad-hoc considerations.

What the manual coding and the empirical comparison of the networks showed is that there is a lot of variability in signatures and devising algorithms able to detect most of it requires a lot of knowledge about the community practices and language (for signature, archiving, use of templates, renaming of users, …) and this is not always feasible. Moreover, it would be unreliable to compare findings across wikis because it could be that the same algorithm performs very differently on different wikis.

On the other hand, the history algorithm, by relying on the edit history produced as a standardly formatted XML file by the Mediawiki software is significantly more robust in each wiki and would also allow comparisons across wikis which are not conditioned by the specific language. Moreover the date of edits, written as standardized timestamps by the platform software, allows to perform longitudinal analyses of the evolution of wikis.

6. CONCLUSIONS AND FUTURE WORK

In this paper we focused on conversations happening on user talk pages of wikis. We proposed to study them from a Social Network Analysis perspective. However we did not enter in considerations about centrality of nodes or topology of the network but we concentrated on the preliminary but essential step of reliably extracting the network. We presented two different algorithms for this task, one looking at signatures in the text of UTPs, the other parsing their edit history.

The contribution of this paper is centered around a detailed comparison of the networks generated by the algorithms running on the Venetian Wikipedia with the real network built by manually coding every message appeared on UTPs. This empirical comparison allowed us to ponder the relative benefits of the algorithms and the issues caused by the inherently free nature of Wikipedia, where rules and practices are continuously negotiated and changed.

The careful comparison showed that the algorithm relying on the history of page edits generated by the platform software is more robust than the one which suffers from the extreme variability of signatures, archiving practices and different languages of wikis. However depending on the hypotheses under analysis, different algorithms with specific ad-hoc heuristics might be appropriate.

We believe that, only by precisely describing how social network extraction is operationalized, it is possible to produce reliable empirical findings. The goal of this work is in fact to start a solid effort for the computational quantitative sociology of wikis as platforms in which millions of people together create resources. To this end, we released the scripts we developed as open source so that other researchers can build on them. We also released some network datasets for other researchers to analyze.

7. ACKNOWLEDGMENTS

Our thanks to Marco Frassoni and Davide Setti who wrote the actual Python code and performed the manual coding.

8. REFERENCES

[1] Giles, J. 2005. Internet encyclopaedias go head to head. Nature, Vol. 438, No. 7070.

[2] List of largest wikis. Retrieved on January 29, 2011 at http://meta.wikimedia.org/wiki/List_of_largest_wikis

[3] Mislove, A., Marcon, M., Gummadi, K. P, Druschel, P., Bhattacharjee, B. 2007. Measurement and analysis of online social networks. 7th ACM conference Internet measurement

[4] Halvey, M. J., Keane, M. T. 2007. Exploring social dynamics in online media sharing. In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA, 1273-1274.

[5] Zelenkauskaite, A. and Massa, P. 2011. Tracing the interpersonal value of Wikipedia community interaction over time. Under review at Wikisym 2011.

[6] Massa, P and Zelenkauskaite, A. 2011. Digital libraries and social Web: Insights from Wikipedia Users' activities. In Proceedings of IADIS Collaborative Technologies 2011.

[7] 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. Physical Review E - Statistical, Nonlinear and Soft Matter Physics, 74(3 Pt 2).

[8] Bellomi, F., Bonato, R. 2005. Network Analysis for Wikipedia. In Proceedings of Wikimania 2005.

[9] Zlatic, V., Bozicevic, M., Stefancic, H. and Domazet, M. 2006. Collaborative web-based encyclopedias as complex networks. Physical Review E, 74:016115.

[10] Zesch, T., Gurevych, I. 2007. Analysis of the Wikipedia Category Graph for NLP Applications. In Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)

[11] Schonhofen, P. 2006. Identifying Document Topics Using the Wikipedia Category Network. In Proceedings of the International Conference on Web Intelligence.

[12] Viegas, R. B., Wattenberg, M., Kriss, J. and van Ham, F. 2007. Talk Before You Type: Coordination in Wikipedia. Proceedings of HICSS '07.

[13] Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. and Suri, S. 2008. Feedback Effects between Similarity and Social Influence in Online Communities. In Proceeding of ACM SIGKDD international conference.

[14] Leskovec, J., Huttenlocher, D. P. and Kleinberg, J. 2010. Governance in Social Media: A Case Study of the Wikipedia Promotion Process. ICWSM conference, AAAI Press.

[15] Kittur, A. and Kraut, R. E. 2010. Beyond Wikipedia: coordination and conflict in online production groups. In Proceedings of the 2010 ACM conference on CSCW.

[16] Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., Smith, M. 2011. Finding social roles in Wikipedia. In Proceedings of the 2011 iConference. ACM, New York, NY, USA, 122-129.

[17] Kimura, M., Saito, K., and Motoda, H. 2009. Blocking links to minimize contamination spread in a social network. ACM Trans. Knowl. Discov. Data 3, 2, Article 9.

[18] Iba, T., Nemoto, K., Peters, B. and Gloor, P. 2009. Analyzing the Creative Editing Behavior of Wikipedia Editors Through Dynamic Social Network Analysis. COINs Collaborative Innovations Networks Conference.

[19] Suh, B., Chi, E. H., Pendleton, B. A., and Kittur. A. 2007. Us vs. them: understanding social dynamics in Wikipedia with revert graph visualizations. IEEE Symposium on Visual Analytics Science and Technology (VAST '07), 163-170.

[20] Brandes, U., Kenis, P., Lerner, J., van Raaij, D. 2009. Network analysis of collaboration structure in Wikipedia. In Proceedings of World Wide Web conference.

[21] Geiger, S. 2010. What is in Control of Wikipedia? Talk at Critical Point of View Conference.