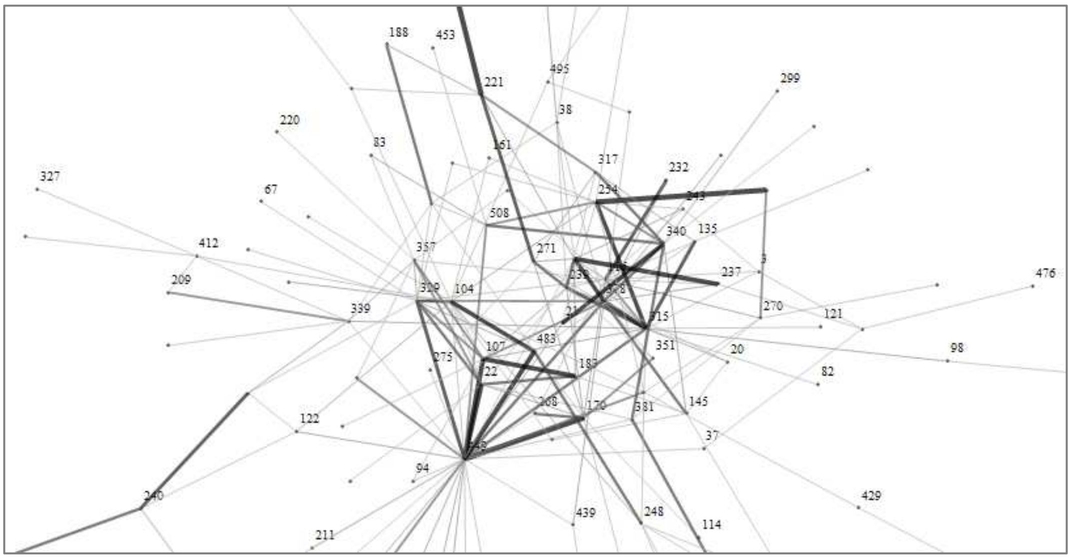
UNIVERSITY OF GOTHENBURG



A Framework for Modeling On-line Communication

Master of Science Thesis in Complex Adaptive Systems

CONSTANTINE A. KULAK

University of Gothenburg
Chalmers University of Technology
Department of Physics
Göteborg, Sweden, December 2012

A Framework for Modeling On-line Communication

Constantine Kulak

Examiner: Dr. Kristian Lindgren

University of Gothenburg
Chalmers University of Technology
Department of Physics
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-786 0000

Cover: A sample fragment of a relationship graph, extracted from the real forum. See section “Acquiring Data from Internet Forum” for details.

Department of Physics
Göteborg, Sweden December 2012

Table of Contents

Introduction 1

Choosing Target Forum 2

Communication Structure 3

Acquiring Data From Internet Forum 3

Structure Recognition Accuracy Experiment 6

Descriptive Statistics 8

Definitions and Method 8

Posting Volume by User 9

Degree Distribution 9

Distance Distribution 10

Reciprocity 11

Volume by Date and Time 11

Posting Rhythm 13

Replies Per Single Message 14

Join and Leave 16

Triad Significance Profile 17

Communication Model 19

Analytical Model 19

Simulating Communication and its Stability 21

Numerical Simulation 23

Emotion Recognition Accuracy Experiment 25

Further Ideas 28

Conclusions 29

Appendix A: Technical Details 30

Appendix B: Modeling Communication Graph 31

References 33

Introduction

Social networks have long been one of the favorite topics of studies for applied mathematicians, and with the recent development of the Intetrnet this area enjoys another wave of popularity. Citation [1] and scientific collaboration [2] networks, email networks [3] and specialized social network websites such as Flickr [4] were thoroughly analyzed, and some of their properties are well known nowadays. Probably the most common approach to the research in this area involves network analysis from the graph theory standpoint, including (but not limited) to calculation of preferential attachment and clustering coefficients, identifying major motifs, and so on. Temporal properties of the networks and their evolution have been studied as well.

The social aspect of the aforementioned networks is often neglected in such research, and the corresponding studies by social scientists traditionally follow a different methodology. Often conducted in restricted artificial settings, those pose a significant challenge for the researcher who wants to make sure the obtained results are correct and unbiased. This is another extreme, as opposed to studying exclusively the network structure and evolution dynamics.

This work aims at outlining some conditions for bridging two approaches together. A relevant sociological data set is obtained from the Internet forum in the automated (and thus unbiased) mode. It is analyzed statistically, and the model of the corresponding processes is created using a specialized agent-based framework. Let us briefly mention few common problems with other research vehicles, which gained popularity primarily due to the data gathering simplicity.

Some websites, such as YouTube or Facebook allow to collect some data related to relationships between the users, but the inherent flaw of such data sources lies in their artificial nature. For example, ability to “like” some content on Facebook, but lack of a symmetric “dislike” feature makes the corresponding study somewhat single-sided. One can argue that studying “dislikes” is even more interesting, because it represents an unusual behavior, as people naturally try to avoid conflicts and negative opinons [5, 6]. Similarly, for the websites a-la LinkedIn the “friends” lists have little relevance to the “real” friends, and new “friends” are recommended by the system itself, significantly affecting the users’ opinions and, consequently, research results. Many of those recommendation algorithms essentially work by “closing triangles” between people, which of course affects the corresponding motif profiles.

Studying other structures, such as email [3] and publishing networks also has its own limitations, because those networks are professional in nature, and thus people do not reveal some important part of their psychology in such interactions. Forums (especially loosely moderated and unspecialized ones), on the other hand, seem to reflect the casual communications rather well. The users demonstrate reciprocal behavior, they gather in groups and form hierarchies, express support and anger towards each other, sometimes going as far as open verbal offense.

Taken that into consideration, Internet forums seem to be promising research targets, however such research is severely impared by the need to analyze the “raw” data somehow. There are no known technical methods able to reliably extract communication semantics, given nothing but the body of text. Furthermore, even human experts not familiar with the “local” rules and trends may experience difficulties trying to interpret some conversations. In this work this task is simplified greatly by considering only the emotional content of the conversation, which seems to be easier to interpret. There is a vast array of possible research opportunities in this framework (some of them are outlined in the final section), and this work covers only few of them. It should be viewes as a source of inspiration for further research, as its primary goal is to outline the methods for analyzing Internet forums and mention some of the early results obtained.

There are four major sections in this work – we start with describing our target and the method for data acquisition. Then some of the “classic” statistics from graph theory are obtained and compared to the similar findings for other well-known social networks. We continue with developing a simple analytical conversation model and analyzing some of its properties. Finally, we come up with a more comprehensive computational model based on the “real world” data, and perform a series of experiments, trying to understand how the users’ relative authorities evolve with time. As a result, we obtain a rather unique framework for further research in social and related sciences.

Choosing Target Forum

Most of the popular forums eventually develop some unique conversation style, a special slang and folklore. As years of communication go by, the users “agree” on some set of rules, often never stated explicitly. Study social dynamics is the primary aim of this work, thus it seems natural to pick some losely moderated forum of the general interest as an experimental vehicle.

After considering a number of alternatives, a choice was made in favour of a Belarusian automotive forum *racing.by* for the following reasons:

1. The author is its active member since August 2009, which allows for early results validation and interpretation.

2. The forum is de-facto unmoderated and the discussion topics range from motorsports to politics. The most popular thread category is “Offtopic”.

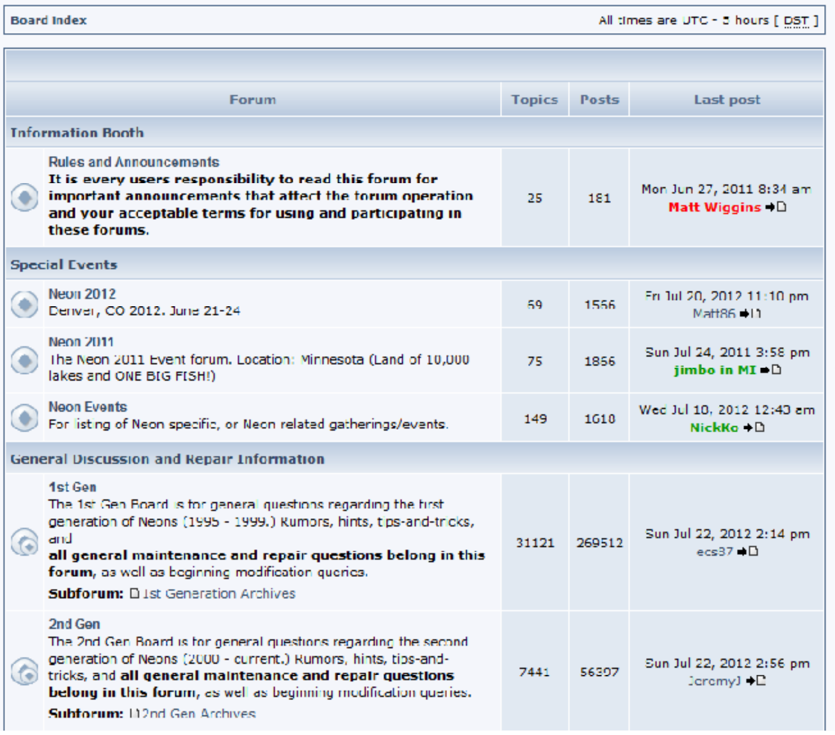
3. The audience consists primarily of young automotive enthusiasts, mostly male. It is a very competitive structured environment, where opinions are questioned and objections are not tolerated. Vulgar language, explicit offence and even direct verbal threat is not unusual. It is definitely not the most welcoming place for inexperienced newcomers.

In this heterogenous ecosystem the users demonstrate wide range of behaviors, which makes this forum a fruitful, yet challenging research target.

Communication Structure

Acquiring Data From Internet Forum

The forum essentially consists of messages (also known as posts), grouped into topics (also known as threads), that in turn are grouped in topic categories (sometimes referred as forums), listed on *board index* page (see Figure 1).

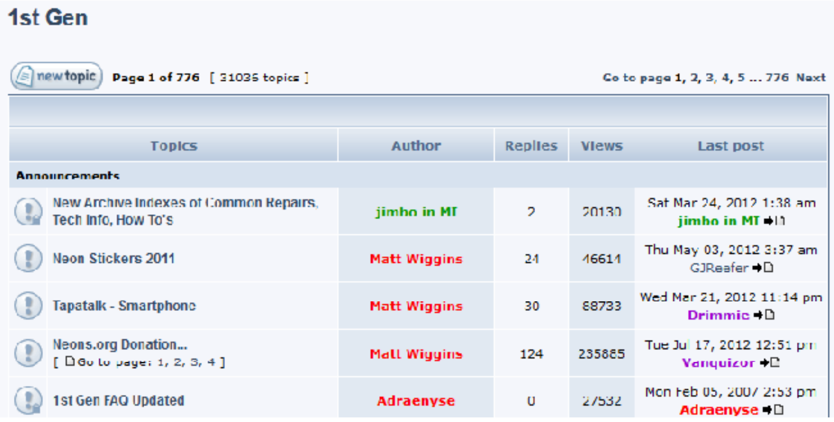


Board Index   All times are UTC - 5 hours [DST]

Forum	Topics	Posts	Last post
Information Booth			
Rules and Announcements **It is every users responsibility to read this forum for important announcements that affect the forum operation and your acceptable terms for using and participating in these forums.**	25	181	Mon Jun 27, 2011 8:34 am Matt Wiggins
Special Events			
Neon 2012 Denver, CO 2012. June 21-24	59	1556	Fri Jul 20, 2012 11:10 pm Matt86
Neon 2011 The Neon 2011 Event forum. Location: Minnesota (Land of 10,000 lakes and ONE BIG FISH!)	75	1856	Sun Jul 24, 2011 3:58 pm jimbo in MI
Neon Events For listing of Neon specific, or Neon related gatherings/events.	149	1610	Wed Jul 10, 2012 12:40 am NickKo
General Discussion and Repair Information			
1st Gen The 1st Gen Board is for general questions regarding the first generation of Neons (1995 - 1999.) Rumors, hints, tips-and-tricks, and **all general maintenance and repair questions belong in this forum,** as well as beginning modification queries. **Subforum:** 1st Generation Archives	31121	269512	Sun Jul 22, 2012 2:14 pm ecs37
2nd Gen The 2nd Gen Board is for general questions regarding the second generation of Neons (2000 - current.) Rumors, hints, tips-and-tricks, and **all general maintenance and repair questions belong in this forum,** as well as beginning modification queries. **Subforum:** 2nd Gen Archives	7441	56307	Sun Jul 22, 2012 2:56 pm JeremyJ

Figure 1. A typical board index. Topic categories, such as "Rules and Announcements" are listed, together with respective descriptions, number of topics, posts and details about the recent activity.

Category page has a similar structure (see Figure 2). Normal users can start new topics, but only forum administrator has enough permissions to alter the categories. There can be thousands of topics in a given category, grouped into pages of 40 topics.



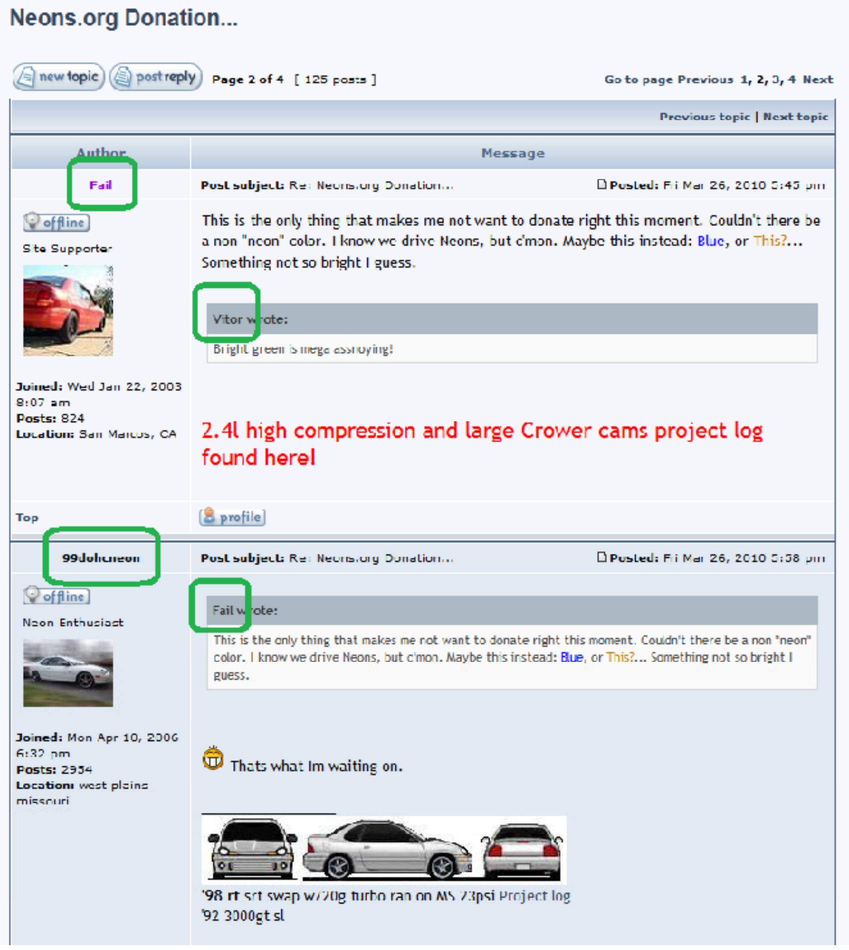
1st Gen

Page 1 of 776 [31036 topics]   Go to page 1, 2, 3, 4, 5 ... 776 Next

Topics	Author	Replies	Views	Last post
Announcements				
New Archive Indexes of Common Repairs, Tech Info, How To's	jimbo in MI	2	20130	Sat Mar 24, 2012 1:38 am jimbo in MI
Neon Stickers 2011	Matt Wiggins	24	46614	Thu May 03, 2012 3:37 am GJReefer
Tapatalk - Smartphone	Matt Wiggins	30	88733	Wed Mar 21, 2012 11:14 pm Drimmic
Neons.org Donation... [Go to page: 1, 2, 3, 4]	Matt Wiggins	124	235865	Tue Jul 17, 2012 12:51 pm Vanquizor
1st Gen FAQ Updated	Adraenyse	0	27532	Mon Feb 05, 2007 2:53 pm Adraenyse

Figure 2. A list of topics on the category page. Note the paging feature and “New topic” button.

Inside a topic the messages are also divided into pages, displaying 40 posts at a time. Apart from the actual message content, each post contains some details about its author and posting time. Figure 3 shows a couple of messages from a typical topic. We are especially interested in quoting feature, which is demonstrated on this figure.



Neons.org Donation...

Page 2 of 4 [125 posts]

Go to page Previous 1, 2, 3, 4 Next

Fail

Site Supporter

Joined: Wed Jan 22, 2003 8:07 am
Posts: 824
Location: San Marcos, CA

Post subject: Re: Neons.org Donation...

Posted: Fri Mar 26, 2010 5:45 pm

This is the only thing that makes me not want to donate right this moment. Couldn't there be a non "neon" color. I know we drive Neons, but c'mon. Maybe this instead: Blue, or This?... Something not so bright I guess.

> Vitor wrote:
>
> Bright green is mega annoying!

2.4l high compression and large Crower cams project log found here!

99dohcneon

Neon Enthusiast

Joined: Mon Apr 10, 2006 6:32 pm
Posts: 2954
Location: west plains missouri

Post subject: Re: Neons.org Donation...

Posted: Fri Mar 26, 2010 5:58 pm

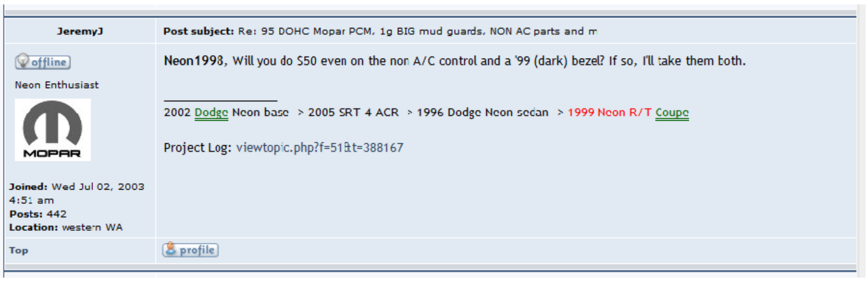
> Fail wrote:
>
> This is the only thing that makes me not want to donate right this moment. Couldn't there be a non "neon" color. I know we drive Neons, but c'mon. Maybe this instead: Blue, or This?... Something not so bright I guess.

Thats what Im waiting on.

'98 rt sct swap w/20g turbo ran on MS 23psi Project log
'92 3000gt sl

Figure 3. A snapshot of two messages in a typical phpBB topic. Note "Post reply" button and paging. User names are highlighted with green, and quotation boxes are clearly visible. By aligning the author name with the quoted content it is possible to identify the exact message the user replies to. In this case it is easy to see that user *99dohcneon* replied to the previous message by *Fail*, about 13 minutes after the latter had posted it. We will refer to this kind of quote as a "full quote", because it allows to reliably identity the source message.

There is another way of replying to others' messages, demonstrated on Figure 4. Here only a user's name is displayed in bold font, and we can only guess what particular message the replier was referring to. Also the message editor allows the authors to use bold font in arbitrary places in their messages, further complicating reliable reply recognition.



JeremyJ

offline

Neon Enthusiast

Joined: Wed Jul 02, 2003 4:51 am
Posts: 442
Location: western WA

Top

Post subject: Re: 95 DOHC Mopar PCM, 1g BIG mud guards, NON AC parts and m

Neon1998, Will you do $50 even on the non A/C control and a '99 (dark) bezel? If so, I'll take them both.

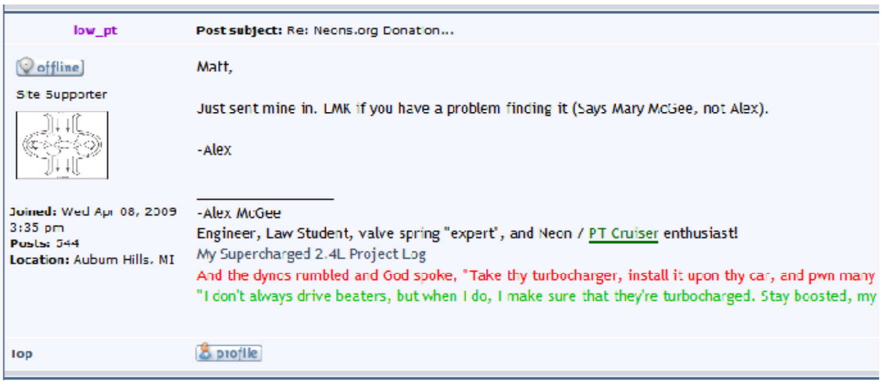
2002 Dodge Neon base > 2005 SRT 4 ACR > 1996 Dodge Neon sedan > 1999 Neon R/T Coupe

Project Log: viewtopic.php?f=51&t=388167

profile

Figure 4. An alternative way of replying. Note the user name (*Neon1998*) written in bold font. Let us call this kind of reply "a simple quote", although technically there is no quote here, but it is implied.

Some users prefer to address the reply target messages informally, for example referring to other users by their real names, instead of nicknames (see Figure 5). In those cases it is especially hard to identify reply target automatically, and thus such messages pose a serious problem for this study. In the following section the severity of this issue is estimated.



low_pt

offline

Site Supporter

Joined: Wed Apr 08, 2009 3:35 pm
Posts: 544
Location: Auburn Hills, MI

Top

Post subject: Re: Neons.org Donation...

Matt,

Just sent mine in. LMK if you have a problem finding it (Says Mary McGee, not Alex).

-Alex

-Alex McGee
Engineer, Law Student, valve spring "expert", and Neon / PT Cruiser enthusiast!
My Supercharged 2.4L Project Log
And the dynos rumbled and God spoke, "Take thy turbocharger, install it upon thy car, and pwn many
"I don't always drive beaters, but when I do, I make sure that they're turbocharged. Stay boosted, my

profile

Figure 5. An "informal" way of replying to another user. A well-known user's real name (*Matt*) is used to address the source message. No special formatting, nor any other hints are given to identify reply target.

Now as the forum's structure is known, let us describe the actual algorithm used for parsing its content:

1. The board index page is parsed, and the list of sub-forums is acquired. Those are identified with some numeric unique identifiers (IDs).
2. For each sub-forum the topic list is parsed. This is a slightly more challenging part, because of the paging mechanism requiring multiple HTTP requests to parse a single sub-forum. Each topic also has its unique ID inside the category.
3. For each topic all messages are parsed by sending corresponding HTTP requests (again, the paging feature has to be taken into account). Messages do not have their own unique IDs, but instead can be identified by their position inside the thread. Therefore, one needs three numbers to address some particular message: sub-forum ID, thread ID and message number inside the thread. The message body is stripped of any HTML markup and quotation content. At the same time, the reply target is identified and is stored as one of the message attributes, together with its author and posting timestamp, which has one minute accuracy.

Identifying reply target is straightforward in case when there is a full quote inside the message, however it is more complicated if we deal with the simple reference to other user's name. Let us consider an example from Figure 4. In this case the algorithm scans up to 50 previous messages in search for the most recent one from *Neon1998* to *JeremyJ,* and if found, it is used as a reply target. If such message is not found, any most recent *Neon1998*'s post will be picked instead. If none of the last 50 messages is posted by *Neon1998*, the algorithm gives up, and the message is not considered a reply anymore. Similarly, if there were neither a quote block, nor some text in bold font in the message body, it is treated as a simple post, and not as a reply.

There is another issue with the quoting mechanism, which is worth mentioning, because it occasionally prevents the algorithm from identifying reply targets. Some users have *automatic censorship* feature turned on (in fact, it is enabled by default for all new users), which automatically replaces some of the curse words in all messages with special text, so that this user cannot see the undesired content. It is important that the user used for parsing the forum has this feature disabled. However this in turn creates an issue: if one poster used a dirty word in his message, and this message was quoted by another poster who had the censorship feature turned on, then this word will not appear in the reply text. In this case the parser will not be able to find the exact match for a quote body among the previous message texts, and thus will have to resort to the “simple quote” method described above.

Some technical details about the parser implementation can be found in Appendix A.

Structure Recognition Accuracy Experiment

In order to estimate the described algorithm’s accuracy, a simple experiment was conducted. A topic with some 250 messages was picked randomly, and a human *expert* analyzed the interactions, deciding which reply target corresponds to each message. Obviously, the sample is small, and the judgment is subjective, but we are doing it only for the rough estimation purposes, thus the accuracy should suffice.

Experiment results are summarized in the table below. Its rows correspond to different users participated in a discussion, thus we can see that, for instance, user 3 posted 22 messages, and the algorithm guessed reply targets correctly for 20 of them (90%), which is a promising result. For other users the algorithm’s efficiency was not as good, scoring 77% on average.

	Posts	Correct	Wrong	% Correct
1	45	35	10	77
2	29	20	9	68
3	22	20	2	90
4	22	18	4	81
5	21	12	9	57
6	13	11	2	84
7	11	9	2	81
8	10	9	1	90
9	10	8	2	80
10	9	4	5	44
11	8	4	4	50
12	7	7	0	100
13	7	6	1	85
14	6	6	0	100
15	4	3	1	75
16	3	3	0	100
17	3	3	0	100
18	2	1	1	50
19	2	1	1	50
20	2	2	0	100
21	1	0	1	0
22	1	1	0	100
23	1	1	0	100
24	1	1	0	100
25	1	1	0	100
26	1	1	0	100
27	1	1	0	100
28	1	1	0	100
29	1	1	0	100
30	1	1	0	100
31	1	1	0	100
32	1	1	0	100
33	1	1	0	100
Total	**249**	**194**	**55**	**77**

It is interesting that these results vary greatly from user to user, probably because of the different individual posting habits. For example, some users always reply by pressing “Quote” button, while others prefer just mentioning person’s name in the message text. It seems that when no clues about the reply target can be found in a message, it is usually safe to assume that the author simply replied to the most recent message. Using this assumption allows to increase the accuracy roughly by 10%, however it was not employed in this study. The “simpler” approach has higher value for scientific studies, because it provides some kind of the bottom margin for the communication volume. In other words, with this algorithm we can be almost sure that the number of identified interactions does not exceed the actual one, which is not necessarily the case with the “guessing” method, which can identify some messages as replies, while in reality they are not.

Descriptive Statistics

Definitions and Method

There are two basic situations when users leave messages on a forum – they either post some information on their own, or participate in communication by replying to other users. In the first case the user can originate a new thread, or write in an existing one. In the second case the user can utilize quotation functionality to specify the reply target precisely. In this case the user has to press “Quote” button corresponding to the message of choice, and the original text of the message is thus included in the reply, together with it’s author nickname. Alternatively the user can click on another user’s nickname, and it will be automatically appended to the current text of reply, without any quoted content however. Finally, the user can always specify the reply target without using forum helpers by simply mentioning another user’s name in the text, or even implicitly when the reply target is clear from the context of conversation. This variety of replying methods poses a challenging problem when one tries to infer the communication order by parsing the forum automatically. It is virtually impossible to implement an easy reliable algorithm for “guessing” the reply target, especially in “implicit” mode.

Let $U = \{A, B, C, \ ...\}$ be the set of registered users that posted at least once. The latter requirement is very important, because it allows to exclude non-existent users from consideration. The parsing algorithm can mistakenly identify text in bold font as a reference to non-existent user, and this condition effectively filters such events out. The filtering has to be done after all data is parsed, because we do not know if the user ever posted something or not, unless we have complete historical record. Anonymous users are unable to post. For convenience let us use small empty circle as a symbol for “no user” and define a corresponding extension for set U: $\overline{U} = U \bigcup \{\circ\}$. Also let’s assume that the users are sorted by descent of the total number of messages they posted during their existence on the forum (with some tie-breaking rule), thus user A wrote more messages than user B, which in turn wrote more messages than C, and so on.

A message $m_i = (a_i, R_i, t_i, s_i)$ has the following components: $a_i \in U$ is the message author, $R_i \in \mathrm{Z}$ is the index of the message a_i replies to (let us use index zero to specify that the user did not reply to anyone, and instead posted the message on his own), t_i is the time stamp of the message in milliseconds since midnight 1st of January 1970, and s_i is the message text (semantic payload). Let us assume that the messages are sorted according to the posting time, i.e. $j \geq k \Leftrightarrow t_j \geq t_k$. Communication happens in the context of K non-intersecting threads $T_j = \left\{m_{j,1}, \ m_{j,2}, \ ..., \ m_{j,N_j}\right\}, \ m_{j,k} \in F, \ T_j \bigcap T_k = \emptyset$, each consisting of N_j messages, and thus the following definitions of the whole forum are equivalent: $F = \{m_i\}_{i=\overline{1,N}} = \bigcup_{j=\overline{1,K}} T_j$.

A method for obtaining communication graphs was outlined above. Such graph from our target forum contains about 860,000 communication events (posts), and it seems natural to analyze it statistically. Different types of social networks are well studied from this point of view, allowing us to compare our results with the published ones.

Before we begin, let us define a key concept that we will use extensively in this study. In order to relate our research to the existing results, we need to obtain a “relationship” graph with nodes corresponding to users, and edges representing some kind of relationship between them. One of the simplest way to do that is to say that the edge from node U_i to node U_j (representing two distinct users) exists only if user U_i wrote at least N_{TR} replies to U_j’s messages. Here N_{TR} is some non-negative constant threshold value that might be useful for skipping “temporary” users that are of little interest in scope of this study. As mentioned above, we can safely drop the users that have some incoming connections, but produced zero messages on their own, because it is highly unlikely that someone will address the user that has never posted anything. The formal definition of the described relationship graph is as following: $\mathrm{Rel} = (V, E)$, where $U \supseteq V = \{v_i \in U\}$ is a set of vertices / nodes, and $E = \left\{\left(v_i, \ v_j\right)\right.$, where $\left|\left\{m_k \in F \mid a_k = v_i \bigwedge a_{R_k} = v_j\right\}\right| > N_{\mathrm{TR}}\Big\}$ contains directed graph edges.

Posting Volume by User

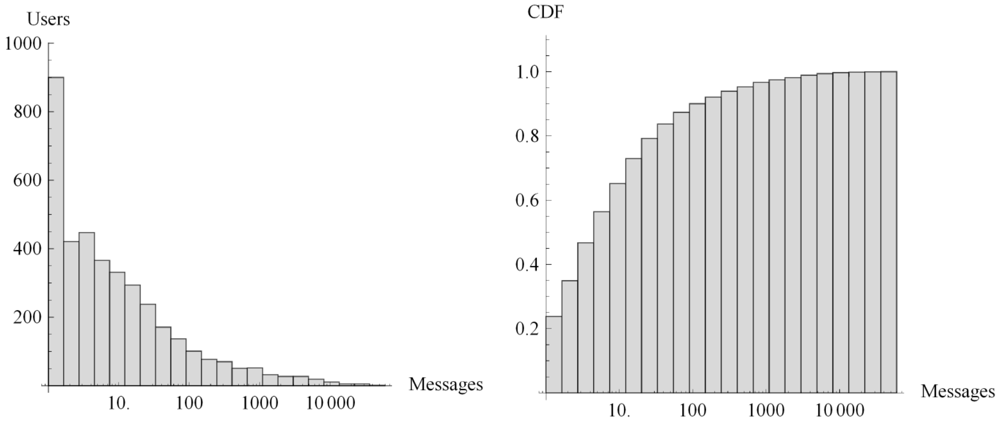


Figure 6. Posting volume distribution. Histogram (left) and CDF (right). Notice logarithmic scale of the messages axis. We can see that 24% of all users wrote only one message, and 65% of all users wrote less than 10 messages. About 50% of the users write less than six messages during their stay on the forum.

Degree Distribution

Let us look at the relationship graph and analyze its' degree distribution. In and out degrees are considered separately, however the statistics are very similar (see Figure 7).

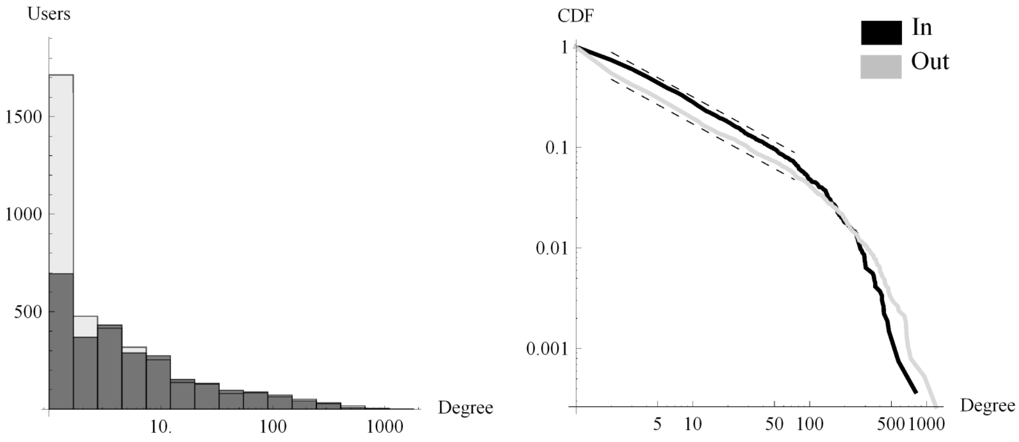


Figure 7. Degree distribution in the relationship graph. Left plot represents the histogram, and the right one demonstrates log-log plot of the cumulative density function (e.g. we can see that about 90% of users have less than 30 outgoing connections).

As expected, the statistics strongly suggest free scaling of the relationship graph, which can be attributed to some variant of the Matthew effect [7, 8].

It might be interesting to look at the similar distribution, if we consider degree divided by volume. In this case the statistics should demonstrate the rate at which the users create new connections. There is an issue with statistics like that, connected to the fact that we plot the histogram of the two integers divided. It is especially noticeable in our case, where the integers are small (e.g. almost half of all users have less than two outgoing connections and post less than five messages in total). As a result, the vast majority of our data points is somewhat granular, with values like 1/1, 1/2, 1/3, 2/2, 2/3, etc. Those are represented as disproportionally high peaks on the histogram. In order to smooth it out, we exclude the users who posted less than 20 (number of bins in our histogram) messages from consideration, thus focusing on "long-term" users (only about 20% of all users).

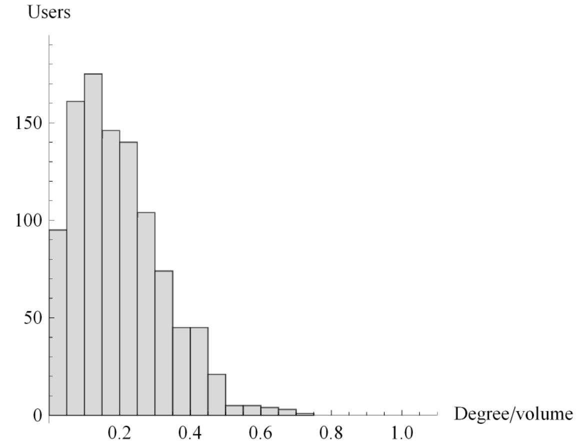


Figure 8. Histogram of the out degree divided by number of messages. Users with at least 20 messages are analyzed in order to reduce the noise caused by the ratio granularity. We can see that on average every 10th message posted by the user is addressed to someone, with whom she has never communicated before.

Distance Distribution

In order to check "small world" hypothesis, let us calculate the distances between the nodes in the relationship graph. Considering only the largest connected component, the shortest distance between each pair of users is calculated, and then plotted in form of the histogram. The results agree well with the findings in other social networks, revealing that 95% of all users are separated from each other with less than four handshakes.

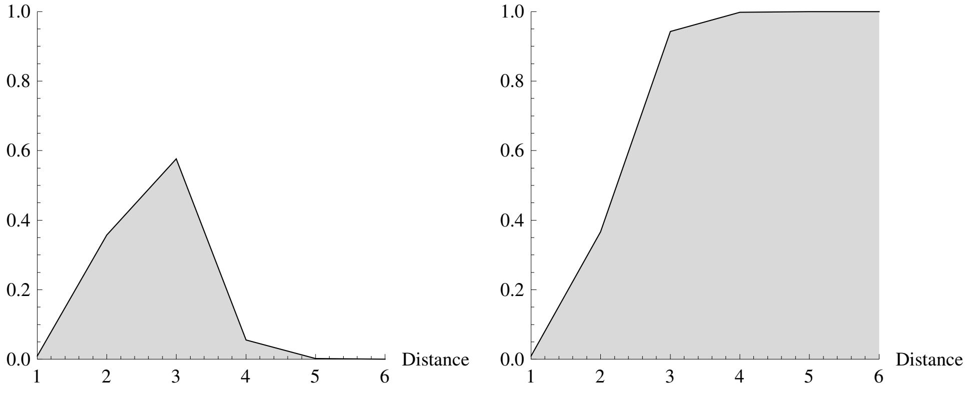


Figure 9. Distance distribution PDF (left) and CDF (right). We can see that 95% of the users are separated from each other with maximum 3 handshakes.

Reciprocity

Let us use the similar approach to analyze communication reciprocity – pick every pair of users that sent at least $N_{\text{interactions}}$ messages to each other, and then compare the number of incoming and outgoing messages within the pair and scale it linearly, so that the value of zero represents no reciprocity (all messages were written in the same direction, e.g. always from A to B or always from B to A), and the value of one corresponds to the situation when both users wrote the same number of messages to each other (fully reciprocal communication).

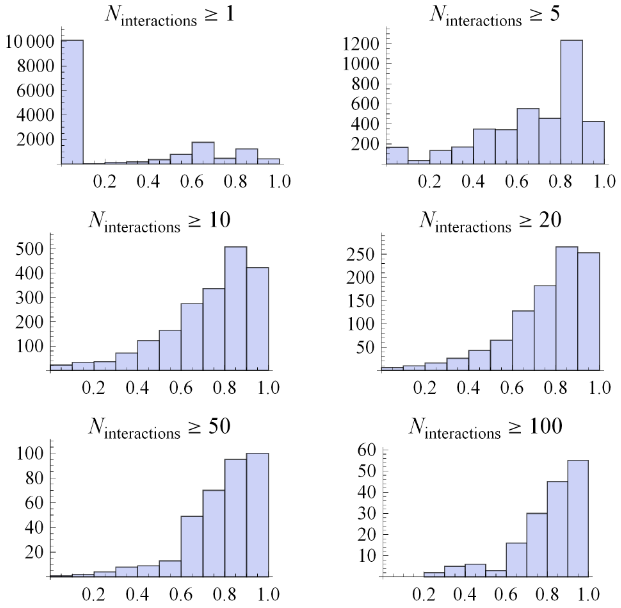


Figure 10. Reciprocity histogram for the users with at least $N_{\text{interactions}}$ messages posted to each other. Reciprocity is measured as normalized difference between sent and received messages for each pair of communicating users, i.e. it is zero if all messages are either outgoing or incoming, and identity in case the number of received messages equals to the number of sent ones. We can see that established relationships tend to be highly reciprocal, similarly to other online social network websites, such as Flickr [4] or LinkedIn [9].

Volume by Date and Time

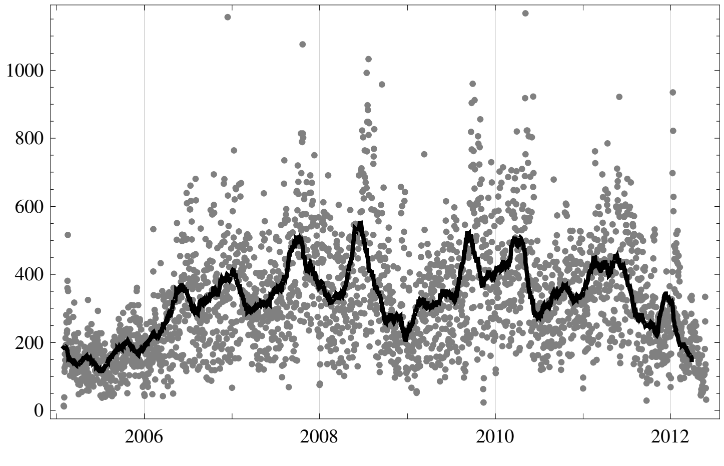


Figure 11. Total posting volume as a function of date (from 2005-01-31 to 2012-05-28, 861,514 messages in total). Black line represents moving average over two months.

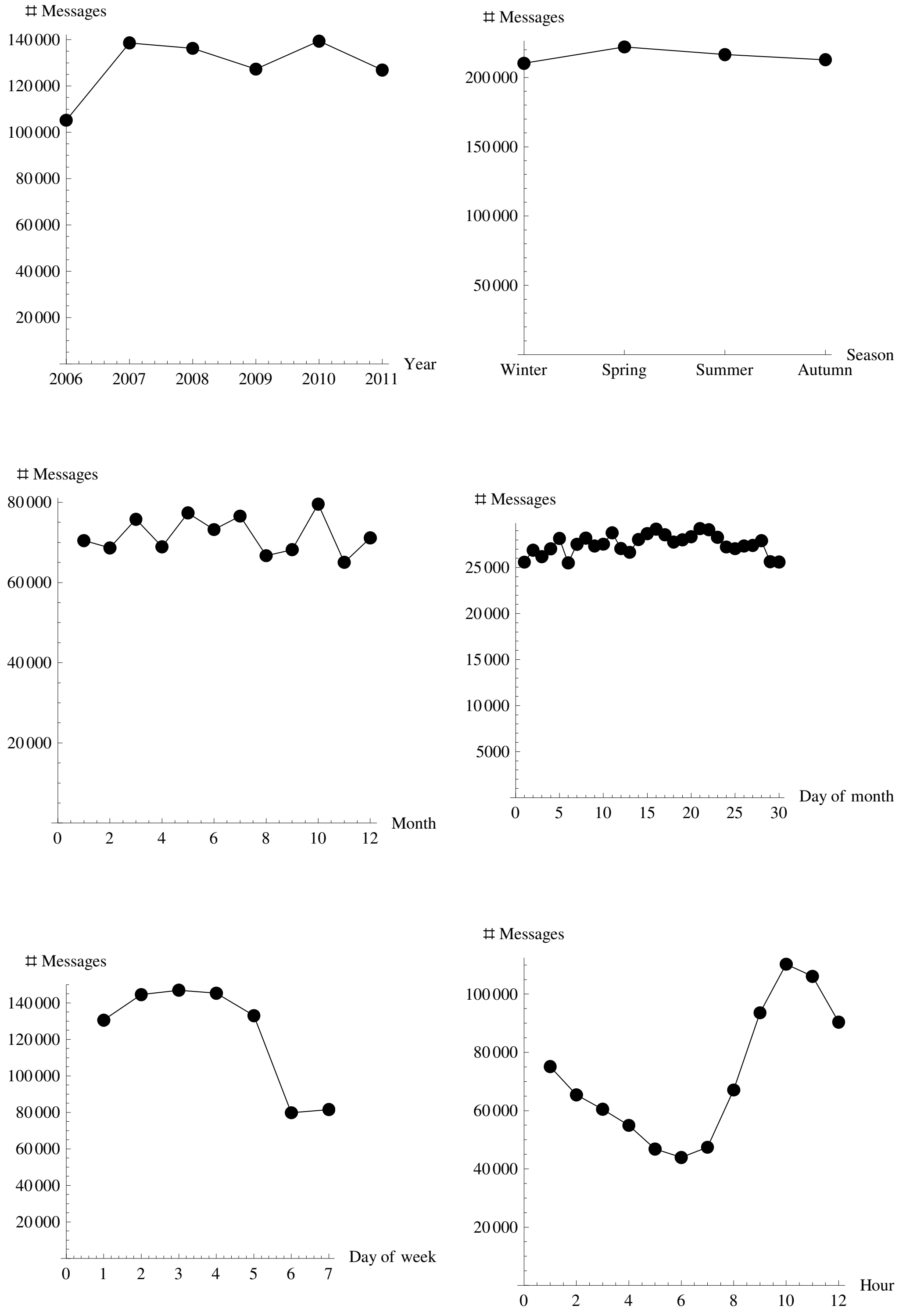


Figure 12. Posting volume by year, month, day of month and day of week (from 2005-02-01 to 2012-01-31). We can see that posting volume fluctuates insignificantly since 2007, and the users are most active in the middle of the week (1 stands for Monday). Posting activity is reduced almost twofold during the weekends. Dependency on the month or season of the year is inconclusive.

Posting Rhythm

When we consider agent-based modelling of communication, it is very important to reconstruct realistic posting sequence. We have already looked at the volume and frequency aspects of the communication, and now it is time to analyze the posting rhythm on the interaction level. Let us take few (frequently posting) users and look at all other users they communicate to. We are interested in timing of the communication events, and more specifically, in delays between each two consecutive events. With statistics like that we shall be able to predict when the next communication event for two given users will happen, based on knowing the last time they communicated.

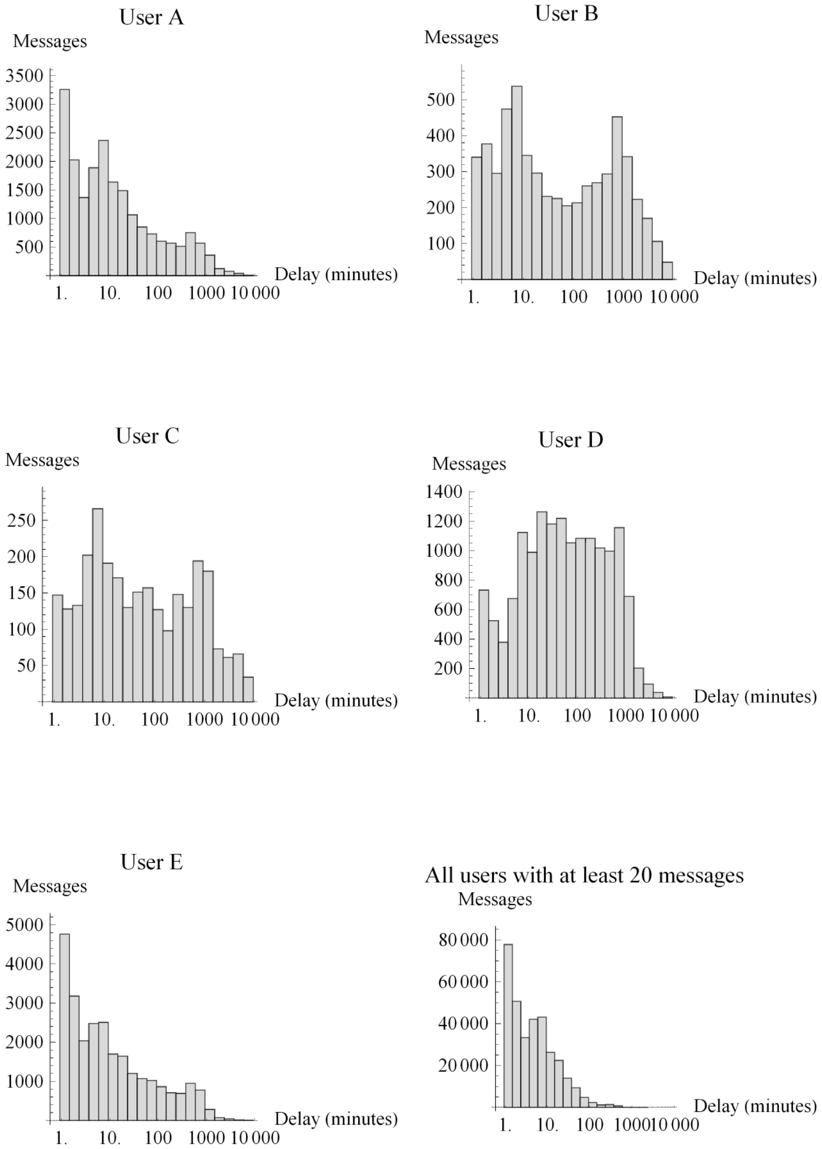


Figure 13. Histogram of posting rhythm for six users with higher posting rates. Delay represents time in minutes between two successive posts. We can see three distinct peaks: around 0 – 1 minute, around 5 – 20 minutes and around 12 – 16 hours. Order and significance of those peaks varies from person to person, implying different posting habits or even amount of thinking time one spends to post a message.

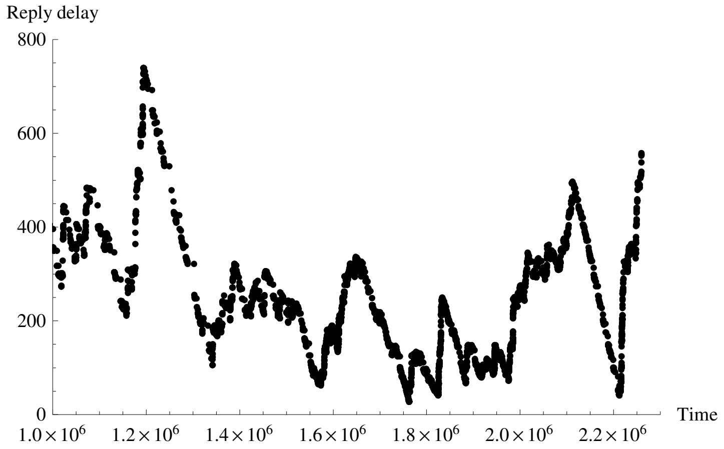


Figure 14. Reply delay for user A (in minutes), smoothed using moving average over 40 observations. Horizontal time scale is in minutes since the first measurement. 0.2×10^6 minutes is ~138 days (4.6 months). We can see that the graph is rather volatile, with regular periods of high activity (approximately three times a year).

Replies Per Single Message

Let us now look at the communication on the single message level. When users write to each other, they form a communication graph (a tree in our case, as we consider every message can reply to at most one other message). Let us take a user and look at the messages she wrote. For each message we know the number of replies it received from other users, and we can obtain the distribution of this number for each user, and then for the whole population. One of the most interesting statistics is the probability of getting zero replies per average message for a given user – in some sense it shows whether others user listen or ignore this one.

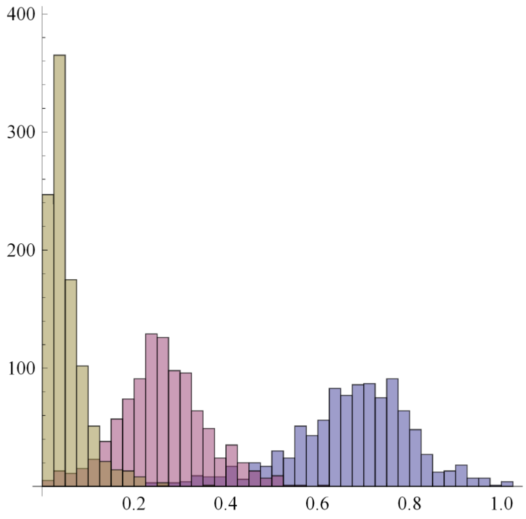
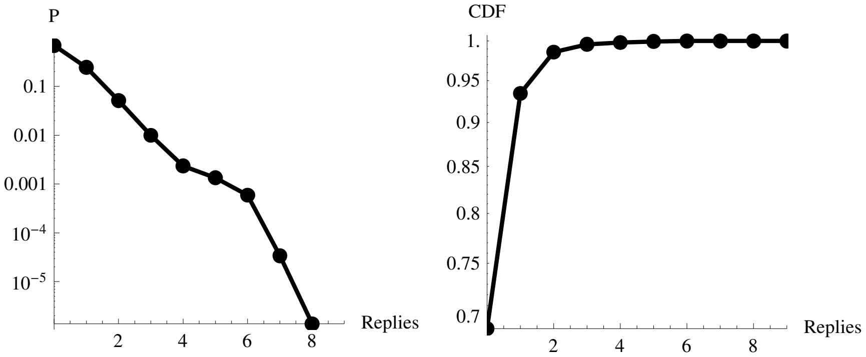


Figure 15. Number of replies distribution histogram for the whole population. Blue chart represents zero replies, purple is for one reply, and the gray is for two replies. Horizontal axis is probability of getting zero, one or two replies, and the numbers on the vertical axis correspond to the number of users with this probability of having X replies. We can see that probability of having zero replies for a given message is approximately twice higher than having one reply, but varies greatly from 0.2 to 1.0, depending on the user.



x	% (replies = x)	% (replies ≤ x)
0	68.9	68.9
1	24.52	93.43
2	5.13	98.57
3	0.99	99.56
4	0.23	99.8
5	0.13	99.93
6	0.05	99.99

Figure 16. Probability density (left, notice log scale) and CDF (right) of the number of replies to an average forum posting. We can see that 99% of the messages have less than four replies, and almost 70% of the messages get no replies at all.

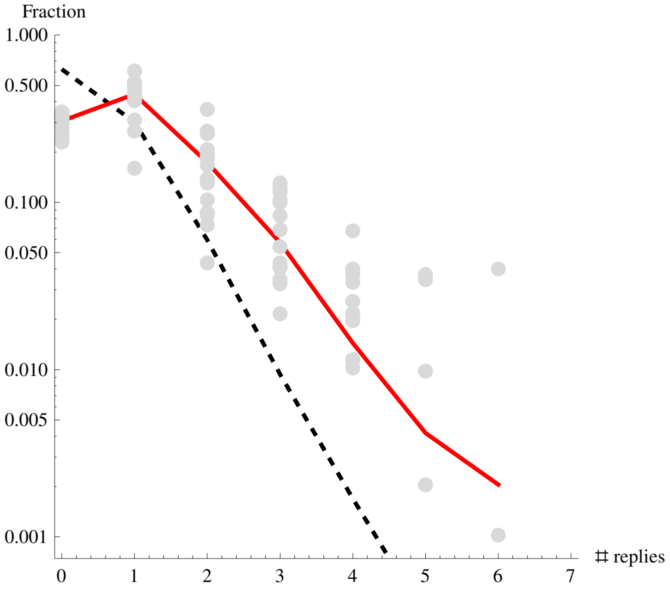


Figure 17. Number of replies distribution for the average user (dashed line, same as on Figure 16), and for the users with unusually high reply rate (red line). We can see that the chance to get no replies is almost twice lower, while the probability of getting two replies is twice as high, and chance to get five replies is ten times higher for those unusual users.

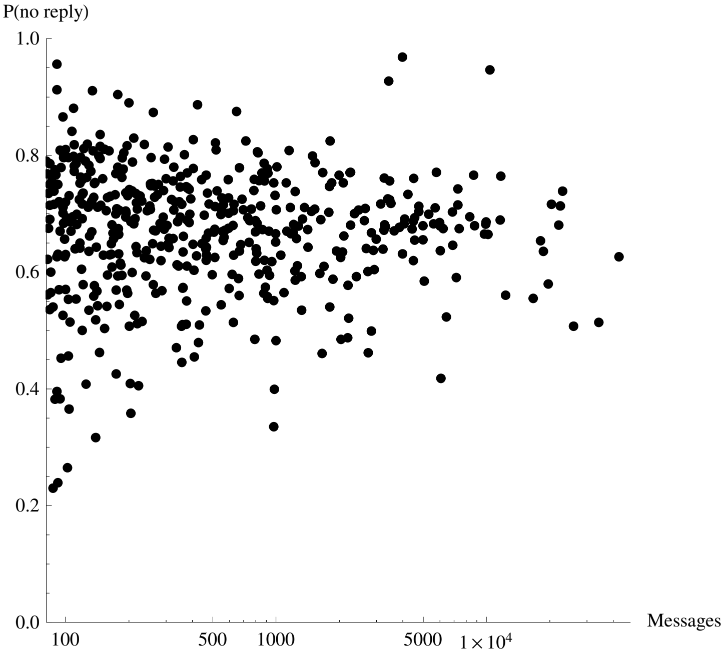


Figure 18. Distribution of the probability of getting zero replies as a function of posting volume (each dot represents a user; notice the logarithmic scale of the horizontal axis). 500 most frequently posting users are plotted. We can clearly see that there are some "outcasts", who post a lot, but receive very few replies. At the same time there is no obvious correlation between posting volume and probability of getting zero replies.

Join and Leave

One of the major drawbacks of the observations above is associated to the fact that we consider complete forum life span, essentially ignoring the temporal aspects of communication. One of such aspects is changing user base – the users constantly join and leave the forum, which in turn affects the posting volume and other statistics. Let us look at the user base size dynamics by analyzing the time stamps of the first and last messages posted by each user, which will give us a rough estimation of their join and leave dates. Let us limit ourselves by considering dates from 1st of January 2006 to 31st of December 2010, otherwise the message time stamps will not reflect the join and leave dates reliably enough.

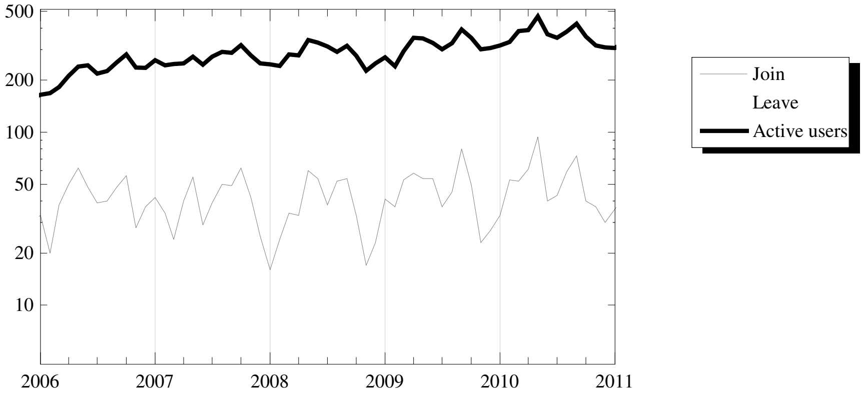


Figure 19. Number of users joining and leaving the forum each month (notice log scale). The plots are correlated because of the short-lived, temporary users with few messages – joining and posting only few times once and never again. However, the overall number of active users is growing steadily.

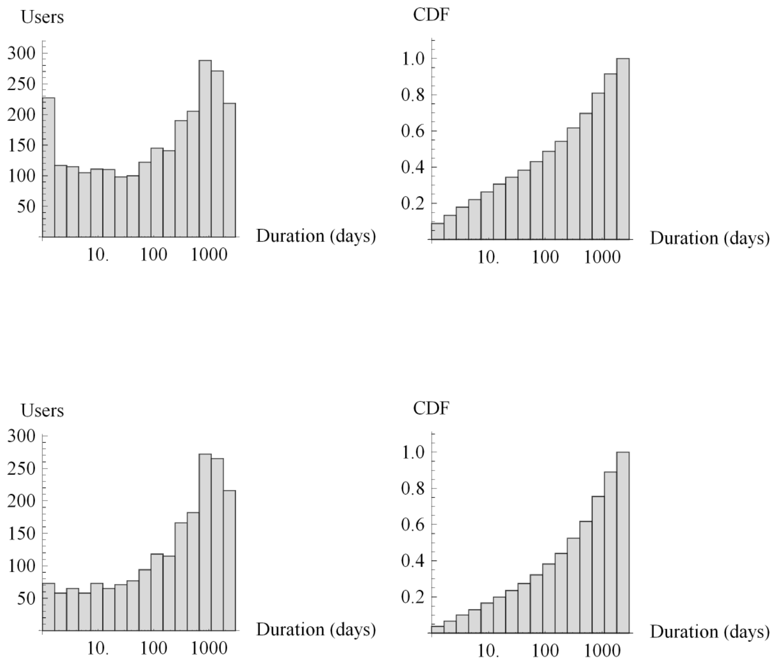


Figure 20. Length of stay on the forum – histogram (left) and CDF (right) for all users (top row) and for those, who wrote at least five messages (bottom row, about half of all users).

Triad Significance Profile

Triad Significance Profile essentially represents statistical significance of non-randomness of triangular subgraphs in the graph under investigation. It is a normalized vector of Z_i scores [10] which are calculated as following: $Z_i = (N\,\mathrm{real}_i - < N\,\mathrm{rand}_i >)/\mathrm{std}(N\,\mathrm{rand}_i)$, where $N\,\mathrm{real}_i$ is the number of times the subgraph appears in the graph under investigation, and $< N\,\mathrm{rand}_i >$ and $\mathrm{std}(N\,\mathrm{rand}_i)$ are the mean and standard deviation of the appearances in the random graph ensemble.

To calculate a TSP for our relationship graph, the edges with less than five interactions in both directions were filtered out. In other words, two nodes representing two users were connected with an edge only if those users wrote at least five messages to each other in total. A relationship graph with 466 nodes and 4621 unidirectional edges is thus obtained.

Values of $N\,\mathrm{real}_i$ were calculated by checking the edges between all possible combinations of three nodes. The task of getting $\{N\,\mathrm{rand}_i\}$ is slightly more challenging, because one has to obtain similar statistics over sufficiently large sample of random graphs, all featuring the same node degrees. Construction of these graphs was handled by the following algorithm:

1. An adjacency matrix representation is obtained for the reference graph.
2. Row i_0 is picked randomly from all rows in the matrix.
3. Column j_0 is picked randomly out of non-zero elements of this row.
4. Row i_1 is picked randomly out of zero elements of column j_0.
5. Elements (i_0, j_0) and (i_1, j_0) are swapped, while the number of ones in column j_0 (in-degree of node j_0) stays the same.
5. The algorithm repeats from step 3, this time using row i_1 as a starting point.

In order to shuffle the edges it is sufficient to iterate this algorithm n^2 times, where n is the number of nodes in the graph. It is easy to see that degrees of the nodes stay the same, except for two nodes, which is acceptable for our purposes.

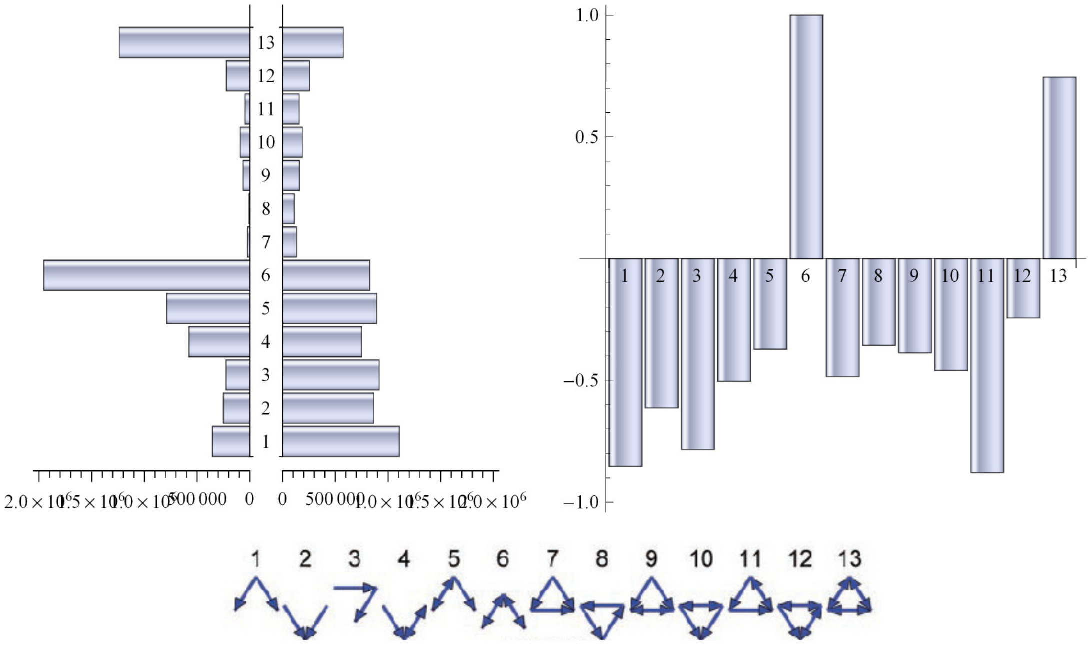


Figure 21. Triad frequency (left) and significance profile (right). Left side of the frequency graph represents real data, while the right one corresponds to the randomized graph with the same degree distribution. We can clearly see that triads with bi-directional links are preferential in relationships graph. Interestingly, this TSP clearly does not fit into any of the superfamilies described in [11], suggesting substantially different nature of this network, compared to commonly studied “friendship” networks [12].

Communication Model

Analytical Model

One of the principal questions this work attempts to address is whether the communication inherently stable or not. In other words, do occasional deviations from the chosen communication strategy dramatically change long-term relationships between the users, or they are tolerated to some extent. This question is essential for this study, because if we observe some chaotic effects of single replies on the long-term authority distribution (i.e. some kind of “butterfly effect”), then with high probability we can claim that communication is practically unpredictable, because interpretation of single messages is very subjective and inherently not reliable. In the following sections we shall see if different reply strategies can damp the fluctuations caused by such misinterpretations.

Let us define a very simple mathematical model for communication and try to study it analytically. In the first section there was given a definition of a message as a vector of author, quotation index, time stamp and text: $m_i = (a_i, R_i, t_i, s_i)$. Imagine some relationship $e(s)$, mapping message text to emotion, quantified as a real value between -1 and 1, where -1 stands for extremely negative emotional content (anger, offense, insult), value of 1 represents very positive emotions (support, compliment), and zero naturally is a neutral message or complete absence of emotions. Let us use the resulting value $e_i = e(s_i) \in [-1, 1]$ as part of the $i-$th message, instead of text s_i. Of course, this kind of mapping is very subjective, as investigated in section 3 later. Right now, however, let us concentrate on our model. For the sake of simplicity, we shall consider the minimal number of users for which we expect some social behavior to emerge – three. Those “users” are very similar – all have the same posting frequency and write to each other by repeating the same posting order every time, effectively producing a *posting cycle* with the period of six (the number of edges in complete directed triangular graph). For the sake of simplicity we can limit the agents’ memory to one posting cycle. Thus, we are dealing with some kind of explicit single-step iterative method over six-dimensional field of parameters $R^6[-1, 1]$.

Also one has to define some order, in which the users will write to each other inside a posting cycle. It is easy to see that particular order is not important, because the users have no access to the information from the current round. For three users A, B and C we can pick the following order: (1: A→B, 2: B→C, 3: C→A, 4: A→C, 5: C→B, 6: B→A). Now we can define some kind of reply strategy for our agents, i.e. some function S of the posting cycle results vector $(a, b, c, d, e, f)_{t+1} = S((a, b, c, d, e, f)_t)$. As for now, let us consider only limited set of deterministic strategies:

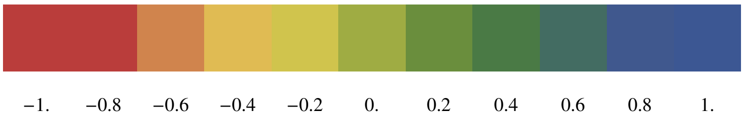
1. *Reply the same as before*: $(a, b, c, d, e, f)_{t+1} = (a, b, c, d, e, f)_t$.

2. *Tit for Tat (TFT)*: when X replies to Y, she writes the same as Y wrote to X in the previous round. For our posting order the strategy can be defined as following: $(a, b, c, d, e, f)_{t+1} = (f, e, d, c, b, a)_t$.

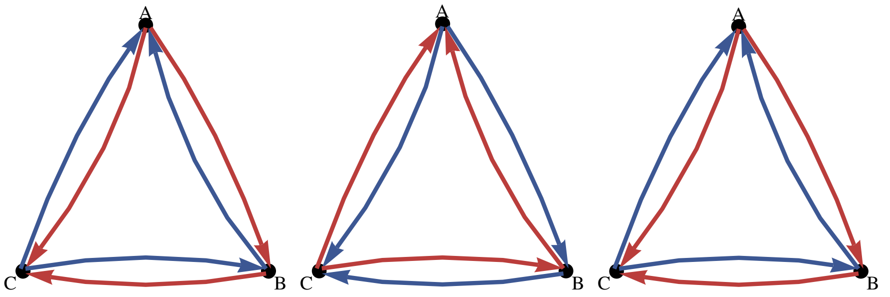
3. *Others’ opinions*: an agent without his own opinion. When X replies to Y she writes what Z replied to Y in the last round: $(a, b, c, d, e, f)_{t+1} = (e, d, f, b, a, c)_t$.

4. *Transitive others’ opinion*: an “improved” version of the previous strategy, when the X takes into account her own opinion about Z when she replies to Y. X’s opinion about Z is reflected by what X wrote to Z on the previous cycle. For simplicity, let us multiply two opinions, so that X writes to Y what Z wrote to Y, multiplied by what X wrote to Z: $(a, b, c, d, e, f)_{t+1} = (e, d, f, b, a, c)_t \times (d, f, e, a, c, b)_t = (e\,d, d\,f, e\,f, a\,b, a\,c, b\,c)_t$.

The communication iterations will be plotted as triangular graphs, using the following color coding for the graph edges, representing posted opinions:

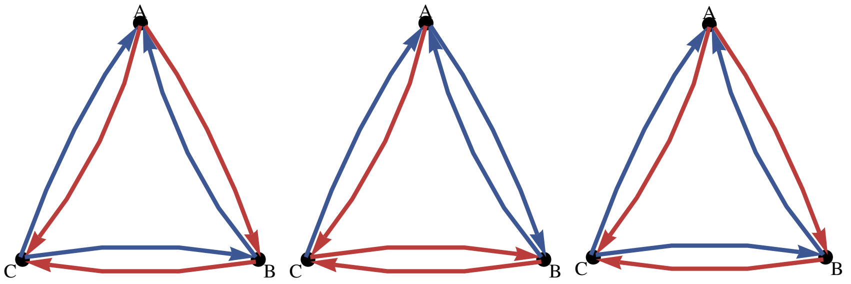


Skipping the first strategy due to its trivial nature, let us take a look at the second one:

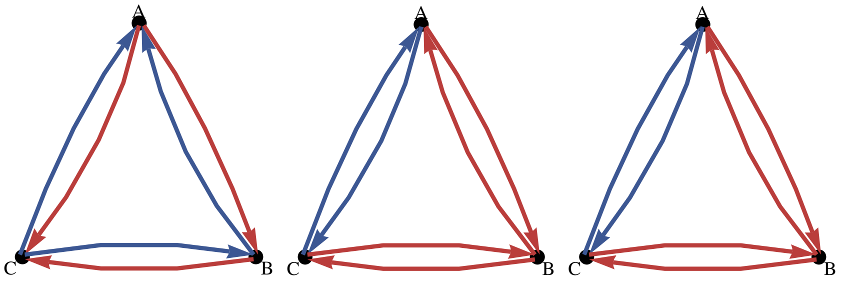


Here a random initial opinion distribution (leftmost graph) is picked, and iterated twice. As we can see, TFT strategy simply "flips" the edges, reverting to the initial state on the second iteration.

The third strategy ("Others' opinions") also has a period of two, flipping the different edges this time:



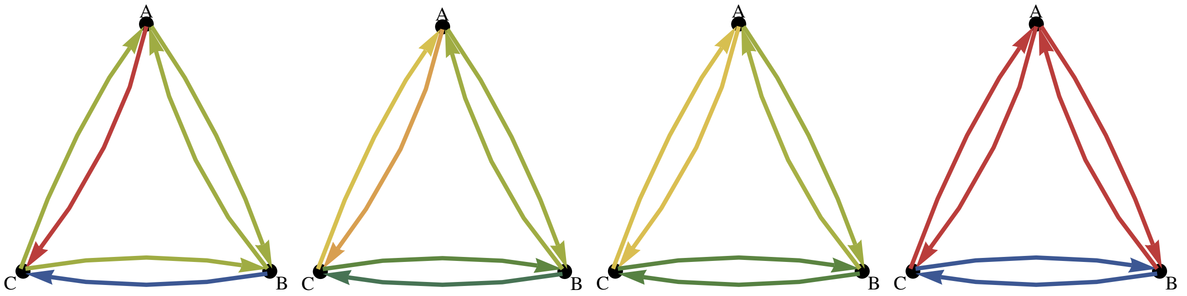
The fourth strategy is irreversible, producing a stable state (we can already see a hint to group formation):



Thus we have four strategies, each one with the distinct features: the first can be considered as some kind of inertia mechanism, preventing the actor from changing opinions; the second (paired with the first one) leads to reciprocal communication; the third strategy amplifies opinions of the majority, and the fourth one leads to group formation. It seems natural trying to combine all four together, introducing four respective weights: (w, x, y, z), $w + x + y + z = 1$, so that our iteration takes the following form:

$$\begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \end{pmatrix}_{t+1} = \begin{pmatrix} a\,w + f\,x + e\,(y + d\,z) \\ b\,w + e\,x + d\,(y + f\,z) \\ c\,w + d\,x + f\,(y + e\,z) \\ d\,w + c\,x + b\,(y + a\,z) \\ e\,w + b\,x + a\,(y + c\,z) \\ f\,w + a\,x + c\,(y + b\,z) \end{pmatrix}_{t}$$

Even with such a simple model we can obtain all kinds of intuitively realistic pictures, such as the one below. Here the following weights were used: (0.5, 0.25, 0.05, 0.2), and initial opinions were distributed such that all actors had neutral relationships, except for user A "dislikes" user C, and B "supports" C. After two iterations we get the situation, where negative and positive opinions become bi-directional, while users A and B develop negative relationships due to conflicting opinions towards C. Signum function is applied to the edges produced on the last iteration to get the rightmost graph.



Now when the communication model is formalized as a map, we can try to analyze its stability, starting by solving the following non-linear system of equations for (a, b, c, d, e, f):

$$\begin{cases} a = (f\,x + e\,(y + d\,z))/(x + y + z), \\ b = (e\,x + d\,(y + f\,z))/(x + y + z), \\ c = (d\,x + f\,(y + e\,z))/(x + y + z), \\ d = (c\,x + b\,(y + a\,z))/(x + y + z), \\ e = (b\,x + a\,(y + c\,z))/(x + y + z), \\ f = (a\,x + c\,(y + b\,z))/(x + y + z). \end{cases}$$

Here the fact $w + x + y + z = 1$ allowed us to exclude weight w from equations. With the explicit expressions obtained for all six variables the system is reduced to three dimensions. Unfortunately, resulting equations have rather complex non-linear form, and thus cannot be solved analytically. However, numerical solution seems possible if weights are specified apriori. Let us further simplify the model by assuming complete communication reciprocity ($a = f$, $b = e$, $c = d$). In this case the system is reduced down to three simple equations, and TFT strategy can be excluded by setting weight x to zero:

$$\begin{cases} a = ((b + c)\,y + 2\,b\,c\,z)/(2\,(y + z)), \\ b = ((a + c)\,y + 2\,a\,c\,z)/(2\,(y + z)), \\ c = ((a + b)\,y + 2\,a\,b\,z)/(2\,(y + z)). \end{cases}$$

This system has five solutions that fall into two groups:

1. Trivial: (0, 0, 0) and (1, 1, 1).
2. “Common enemy”: (X, X, Y), (X, Y, X) and (Y, X, X), where $X = -1 - \frac{3\,y}{2\,z}$ and $Y = 1 + \frac{y}{4}\left(\frac{3}{z} + \frac{1}{y+z}\right)$.

It is easy to see that solutions of the second type are valid (satisfy requirement $-1 \le v \le 1$, where $v \in \{X, Y\}$) only when $y = 0$, i.e. when the actors do not use the third strategy (“unbiased public opinion”). Interestingly, our trivial model suggests that the “common enemy” $(-1, -1, 1)$ configuration is stable, while the “common lover” $(1, 1, -1)$ is not.

Simulating Communication and its Stability

Although being extremely simplified, our analytical model suggests that at least in some cases we can expect communication to be stable, so that small errors in replies are damped, or at least do not result in dramatic changes in long-term authority distribution. Let us extend our model and perform a numerical simulation to further validate this assumption. Its’ core feature is use of the real communication graph, acquired from our target forum. This gives us an important advantage of having correct statistics, so at least we do not have to worry about modeling communication sequence, which is a non-trivial task of its own. Instead, we can concentrate on simulating the agent behavior on the microscopic level, and then compare the emerging statistics to draw some conclusions about the underlying dynamics. Similar approach is used in [9] to validate some hypotheses about the growth mechanisms behind various social networks.

The basic procedure is as following: a communication graph from some relatively long period of time is split into two parts, and some emotional weights (positive, negative or neutral) are assigned randomly to all posts from the first part. Then the agent-based simulation of communication is performed for the second part. Finally, the distribution or authority between the users is calculated and saved for further comparison. The simulation is ran a number of times, varying some parameters of the model, and resulting authority distributions are compared.

Before we dive deeper into the model details, let us define *authority* in a strict mathematical way. One of the possible ways to tell what user A thinks of user B is to look at their communication history, and more precisely at what A ever wrote to B. Relationships evolve with time, thus recent communication events have higher significance. Here is one of the possible formulas to express authority of j_{th} user in the eyes of i_{th} user at time t (in other words, what user i thinks of user j at τ_{th} iteration): $a_{i,j}(\tau) = \sum_{k=0}^{\tau} \delta_{i,j,k}\, e_k\, \lambda^{t_k - t_\tau}$, where $\delta_{i,j,k} = 1$ if k_{th} message is i_{th} reply to j, and zero otherwise, so that the summation goes only over messages written by user i to user j. Emotion $e_k \in \{-1, 0, 1\}$ was defined in the previous section; t_k is k_{th} message timestamp in days since some fixed date; $\lambda = 2^{1/\text{EHL}}$ is some *emotion decay* exponent, where EHL stands for *emotion half life* in days, which is voluntarily taken to be 30 in this work. See Figure 22 for some examples of authority $a_{i,j}$ as a function of time. Note that values of $a_{i,j}(t)$ are not normalized in any way and thus should not be used in computations directly. This simple definition of authority gives us an *authority matrix* $A(\tau) = \{a_{i,j}(\tau)\}$ at every time τ, which will be used extensively throughout the rest of this study.

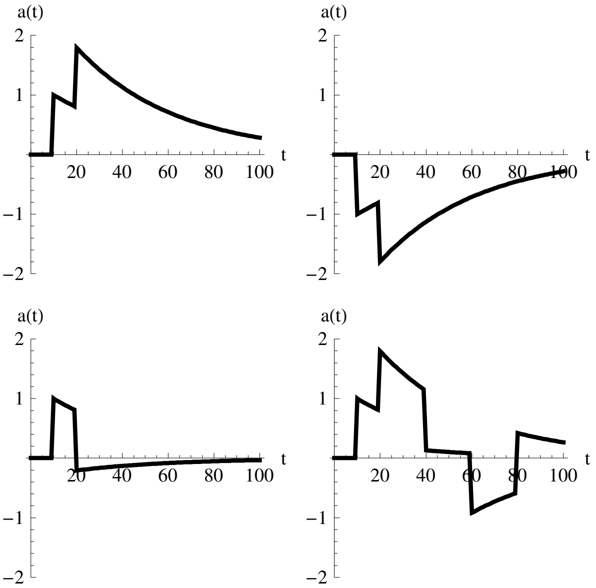


Figure 22. Examples of authority $a(t)$ between two users as function of time t (in days). Events are (top to bottom, left to right): Positive messages at $t = 10$ and $t = 20$; Negative messages at $t = 10$ and $t = 20$; Positive message at $t = 10$, negative message at $t = 20$; Positive at $t = 10$, $t = 20$, $t = 80$, negative at $t = 40$, $t = 60$;

Now let us extend the strategies defined in the previous section to a more realistic setting. The main difference is that now we deal with arbitrary number of users posting messages in random order. As we cannot rely on the fixed iterative structure of communication anymore, some equivalent of the “previous iteration” concept has to be defined, and after modification the strategies for user A writing to B look as following:

1. *Reply the same as before*: iterate through interactions backwards in time, looking for A writing to B. For the most recent event of this type calculate its “value” by using the same *authority decay* function as above, only without the summation.

2. *Tit for Tat (TFT)*: same as the first one, but looking for B writing to A.

3. *Others’ opinions*: looking for users other than A writing to B before. For each user writing to B take the most recent message and calculate its “value” as above, depending on time passed since the event. After all interactions are processed, compute the average for all such “values” and use it as “other’s opinion”.

4. *Transitive others' opinion*: the same as the previous strategy, but the average is weighted with A's "opinion" about the corresponding users, calculated in the same way as in the first strategy.

As in the analytical model, results obtained from all four strategies are weighted with coefficients ($w,\ x,\ y,\ z$), corresponding to strategies 1, 2, 3 and 4, respectively.

■ Example

This model is used in the next section, thus it is essential to understand how it works. Let us demonstrate is on a simple example. Consider the following sequence of events:

⌗	When	From	To	Emotion
1	2012 / 01 / 01	A	B	1
2	2012 / 01 / 02	B	A	1
3	2012 / 01 / 10	C	B	-1
4	2012 / 01 / 20	C	D	1
5	2012 / 02 / 10	D	B	1
6	2012 / 02 / 20	B	A	-1
7	2012 / 03 / 01	C	A	1
8	2012 / 03 / 05	A	D	-1
9	2012 / 03 / 10	A	B	?

Let us define strategy weights as (0.6, 0.2, 0.05, 0.15) and calculate what user A should reply to B in the given situation on step 9:

1. Reply the same as before: A already wrote a positive message to B on step 1, thus $e_1 = 1\,\lambda^{-69} \approx 0.203$, where 69 is the number of days between events 9 and 1.
2. Tit for Tat (TFT): B wrote to A twice – at steps 2 and 6. We pick the most recent event and perform a similar calculation: $e_2 = -1\,\lambda^{-19} \approx -0.645$.
3. Others' opinions: we are looking for events when other users wrote to B. There are two events of this kind: 3 and 5. For each of them we calculate the actual emotion as before: $e_{31} = -\lambda^{-60} \approx -0.25$, $e_{32} = \lambda^{-29} \approx 0.512$, and then take the average value: $e_3 = (e_{31} + e_{32})/2 \approx 0.131$.
4. Transitive others' opinion: the same as above, but we take the weighted average. First find all events where A wrote to C or D and calculate respective emotions. Only event 8 satisfies this criterion, and it gives weight $w_{42} = -\lambda^{-5} \approx -0.891$, which results in the following value for transitive opinion: $e_4 = e_{31}\,w_{41} + e_{32}\,w_{42} = e_{32}\,w_{42} \approx -0.456$.

Weighting the values from separate "pure" strategies, the final "mixed" result for reply is calculated as following: $e = 0.6\,e_1 + 0.2\,e_2 + 0.05\,e_3 + 0.15\,e_4 \approx -0.069$. Now we need to decide what A has to write on step 9, based on this value. One can think of different methods for making such a decision, but to avoid unneccessary complexity this study uses the simplest threshold-based approach: user A will write to user B neutrally, if $|\,e\,| \leq \sigma$, positively if $e > \sigma$, and negatively if $e < -\sigma$, where σ is some fixed threshold value. In this work the value of σ is set to 0.1, as it results in a rather realistic distribution of positive ($\sim$20 %), negative ($\sim$20 %) and neutral ($\sim$60 %) messages. In the example above $|\,e\,| < \sigma$, thus A will write neutrally to B on step 9.

Numerical Simulation

Now with an agent-based simulation model available, let us conduct a series of experiments to outline some conditions under which it seems to produce *stable* results. In order to minimize any "boundary" effects, let us pick the time frame somewhere from the "middle" of the data set – six months from 1st of January 2010 to 1st of June 2010. This interval contains 67 558 messages written by 1246 users. 33 690 or those messages were identified as replies, providing a realistic communication sequence.

The emotional payload is assigned randomly for the first 20 000 messages (roughly two month period from 1st of January 2010 to 5th of March 2010). It is saved and then reused in all experiments. The simulation starts from the message 20 001, calculating the emotions for each message till 40 001, skipping those without identified recipients.

Let us start with completely deterministic model, where all agents reply using the same strategy with fixed constant weights and there is no random noise of any sort. The authority matrix is computed on every 1000_{th} iteration, it gives us the *reference authority distribution*, consisting of a set of authority matrices $\{A_{\text{ref}}(t)\}_{t\in\{20\,001,21\,001,\ldots,40\,001\}}$. Then some parameters of the model are altered, and the simulation is ran again from the start (iteration 20 001) to obtain another authority distribution $\left\{A_{\text{exp}}(t)\right\}_{t\in\{20\,001,21\,001,\ldots,40\,001\}}$.

The absolute authority values are not normalized, and therefore do not make much sense per se. One of the possible methods to compare authority distributions in this situation is to translate real values $a_{i,j}(t)$ into familiar "emotion" form $e_{i,j}(t) \in \{-1, 0, 1\}$ using some empirical threshold value τ: $e_{i,j}(t) = 1$ if $a_{i,j}(t) > \tau$, $e_{i,j}(t) = -1$ if $a_{i,j}(t) < -\tau$ and $e_{i,j}(t) = 0$ otherwise, where $\tau \equiv 1$ is constant, selected to allow fair number of ones and zeros in authority matrices. Now comparing two authority distributions is easy – we can just take a simple Euclidean norm of the difference: $d(t) = \left\|E_{\text{exp}}(t) - E_{\text{ref}}(t)\right\|$ ("*authority difference*").

Unfortunately, it is still hard to give any reasonable explanation to particular values of $d(t)$, unless we can compare them to something. One of the candidates is function $d_{\text{rand}}(t) = \|E_{\text{rand}}(t) - E_{\text{ref}}(t)\|$, where E_{rand} is a simplified authority distribution obtained as a result of randomizing replies 20 001 to 40 001, instead of applying our agent-based simulation. The rationale behind it is simple – if our experimental authority difference $d(t)$ is close to $d_{\text{rand}}(t)$, obtained from completely random results, it means that the simulation has essentially failed, as it results in a random authority distribution. If, on the other hand, experimental authority difference is close to zero, it means that the simulation results resemble the reference, i.e. the algorithm converges to some solution. In order to keep it as objective as possible, the distribution of emotions (percentage of negative and positive responses) for randomized replies is kept the same as for the non-randomized ones.

The model implementation was validated using the above example to minimize the risk of coding mistakes.

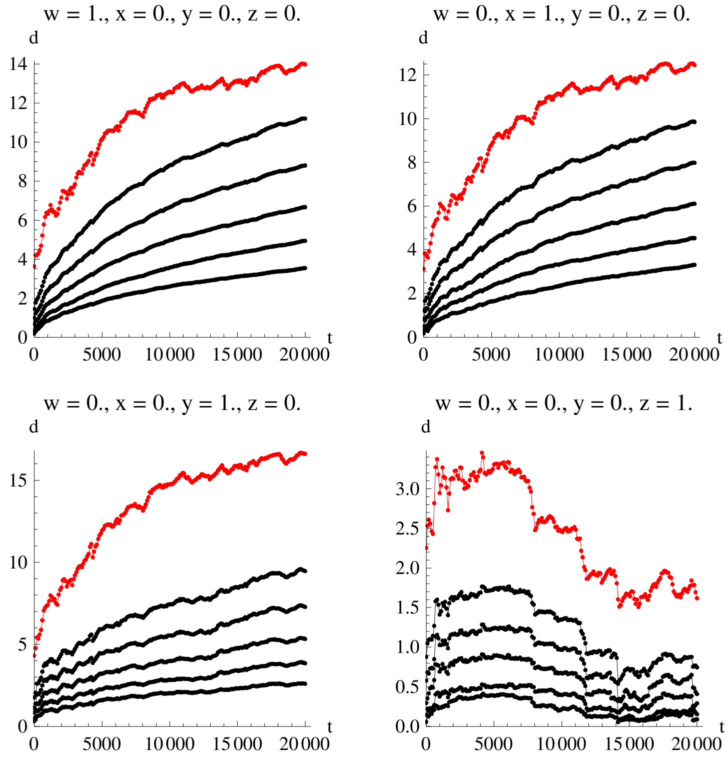


Figure 23. "Pure" strategies, the graphs show the simulation results for error ratios of (bottom to top) 1 %, 2 %, 4 %, 8 % and 16 %. Red plot corresponds to 100 % error ratio (completely random guesses for comparison). The experiment was repeated 100 times to smooth the plots. We can see that pure "reply the same as before" and "tit-for-tat" strategies behave very similarly, and the other two "collective" strategies demonstrate different behavior, especially "transitive others' opinion", which in fact is the only truly stable one, seeming to converge to some kind of stationary state.

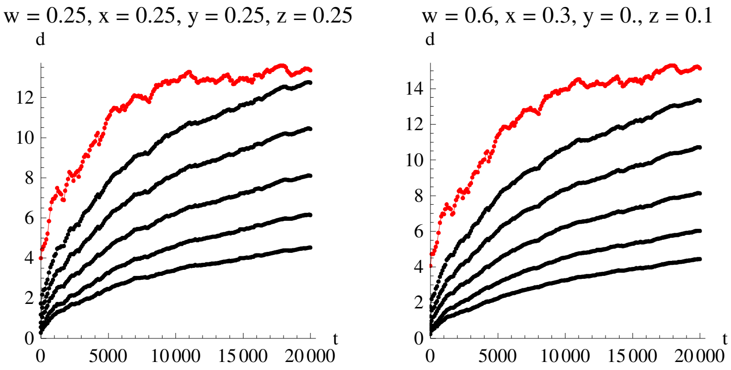


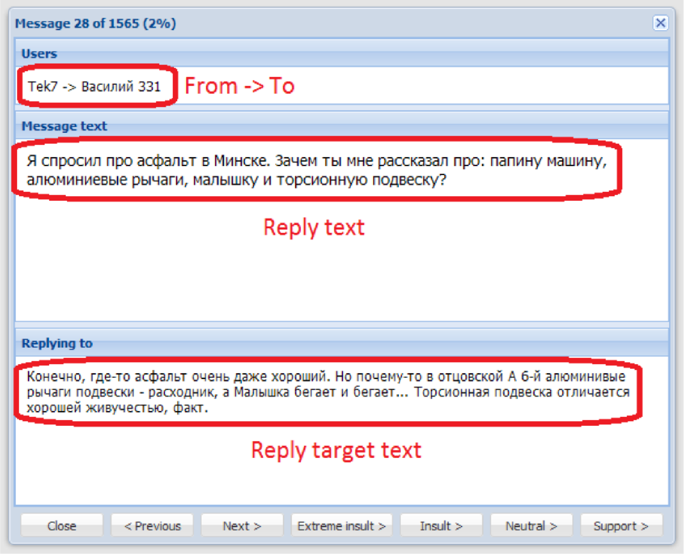
Figure 24. “Mixed” strategies do not seem to perform as well as the “pure” ones, resulting in larger errors, and consequently less predictive power. However, mixing those strategies has to be done with greater care, because their components were not normalized in the first place. Therefore weighting essentially normalizes the values of the components, and thus should be done after analyzing their respective impact on the sum.

Figures 23 and 24 shed some light on the prediction dynamics. Indeed, we can compare $d(t)$ values with different fractions of “noise” against the completely random ones (red plot corresponding to $d_{\text{rand}}(t)$). For example, let us consider left part of Figure 24. It can be interpreted as follows: if our model is iterated and 16 % of its replies are modified randomly, then after 20 000 iterations the authority distribution will be pretty much random as well. However, if we look at the case $w = 0$, $x = 0$, $y = 1$, $z = 0$, we can see that this strategy behaves much better in the same situation, still providing essentially non-random results for authority distribution. Pure “transitive others’ opinion” strategy is probably the most interesting one, because unlike the other three it seems to converge. Even for non-converging strategies we can see that the difference of authority distributions as a function of time does not change dramatically as we introduce more noise, in other words the errors we receive at the output depend on the measurement errors linearly or even sub-linearly. Nothing like the exponential growth of errors, nor any kind of chaotic behavior were ever observed. This implies that at least small changes in input parameters (*measurement noise* in our case) result in relatively small changes in the output (authority distribution), although it will not be ignored in the long run (the solution does not seem to converge in most of the cases).

Emotion Recognition Accuracy Experiment

In the previous section we saw that for the simple communication models some degree of linear dependency between authority distribution and emotional content of the replies looks like a reasonable assumption to make. This linearity is somewhat intuitive as well, because some messages could be misinterpreted or understood differently in different contexts. This section aims at estimating such subjective component of the communication.

In order to quantify the differences in people’s judgments a simple experiment was conducted. For the 125 first messages from two rather “emotional” topics a group of seven people was asked to rate each of those 250 messages as either supportive, neutral (or unknown), insulting or extremely insulting. In order to simplify the experiment, a web application was implemented and hosted on the Internet (see Figure 25). Only one participant (user A, the author of this study) was familiar with the forum and context of conversation. Respondents’ age varied from 23 to 28 years, all of them lived in the capital city and had higher education. Six out of seven were males, and five out of seven had driving licenses. This sample seems to be rather representative, taking into account the forum’s topic, typical audience and general sociological context.



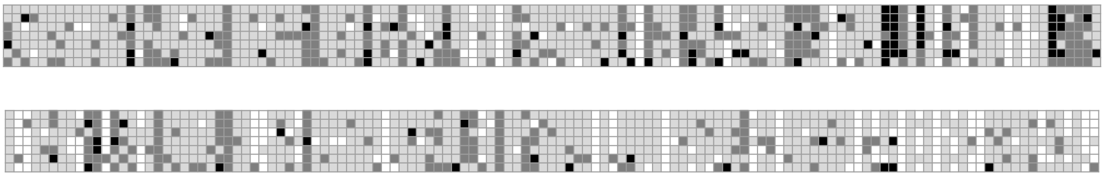
Tek7 -> Василий 331

Я спросил про асфальт в Минске. Зачем ты мне рассказал про: папину машину, алюминиевые рычаги, малышку и торсионную подвеску?

Конечно, где-то асфальт очень даже хороший. Но почему-то в отцовской А 6-й алюминивые рычаги подвески - расходник, а Малышка бегает и бегает... Торсионная подвеска отличается хорошей живучестью, факт.

Figure 25. A sample screenshot of the emotion rating web application. For each message in the thread the users see its author and its text (converted to plain text). If the algorithm detected a quote, then the user also receives information about the corresponding reply target. A set of buttons for rating emotions is available aside the navigation controls (the user can skip the messages, for example in case of uncertainty or ambiguity).

Figure 26 demonstrates the survey results. We can see that for some posts all respondents gave the same ratings, while for the others their opinions were more diverse. Five out of 250 messages received both “supportive” and “extremely insulting” ratings, and 25 were classified as “supportive” and “insulting” by different people, which demonstrates perception subjectiveness and implies the importance of understanding the communication context. Survey data was compared to the reference (number of different replies was counted), which provided a rough estimation of the subjectiveness – about 77 % of the reference rankings were “correct”, i.e. shared by majority of respondents.



	A	*B*	*C*	*D*	*E*	*F*	*G*	Mean
$\%_{\mathrm{diff}(A)}$	–	16.8	24.8	27.6	14.4	23.2	31.6	**23.1**

Figure 26. Results of the experiment. The narrow rows correspond to each of the seven survey participants, while the columns represent the messages in the topic (125 consecutive posts in two topics were analyzed). The table summarizes differences in opinions, compared to the user A (author). We can see that on average 76.9 % of A’s opinions were shared by the rest of the participants.

This experiment suggests that the error rate of evaluating the emotions is on the order of 20 %, which renders our model barely useful. In the previous section we observed very little to none long-term predictive power for comparable amount of errors, thus it cannot be used for reliable prediction of authority distribution, at least in the form it was described and implemented. On the other hand, even the simplest model (where agents' memory was limited to a single event, and all of them shared the same communication strategy) provided substantial degree of non-randomness in short-term authority prediction (up to about one week). It leaves us some hope that if the model is developed further (for example, the agents receive more memory and their strategies are weighted individually), it could result in sufficient performance improvement.

Further Ideas

It appears that potential scope of the studies in the area under investigation is remarkably broad, which is not surprising taking into account the origin and amount of data at researcher's disposal. This work merely scratched the surface of the problem, aiming at establishing a basis for further advancement in this area. Let us glimpse at few examples of the directions for potential research.

1. Use some keyword-based semantic analyzer to parse emotional data from the messages automatically. One can immidiately see multiple potential problems with this approach – for example it might detect insults better than positive responses, simply because the latters are usually not expressed as explicitly, and therefore the research will be effectively single-sided (similarly to analyzing Facebook-like networks, where we have only the information about some positive reactions in form of "likes"). Although it is expected to provide only a fraction of data, this could be enough for giving some predictions, taking into account the large overal size of the data sample. In fact, the keyword-based recognition was implemented in the parser, however it has never been configured for the real forum, thus its performance has not been tested.

2. The parser implementation is generic and should work for any phpBB forum, which means analyzing different forums and comparing resulting statistics are both relatively straightforward tasks. There are thousands of phpBB forums all over the Internet, and therefore such research could potentially reveal some interesting differences and correlations, related to culture, language, geography, topic or size of the forums, just to name a few.

3. If we had large sample of semantic (emotional) data from the real forum, then it should be possible to solve the inverse problem, i.e. compute parameters for our agents using some optimization (e.g. evolutionary) algorithm [13]. For example, the chromosomes consisting of strategy weights (w, x, y, z) could be obtained for each agent using the genetic algorithm with authority distribution error as its fitness function. When all such parameters are known, we can test if the model has some real predictive power. In this case it could be used as some kind of a risk assessment tool to model the possible outcomes of user actions. Of course, if such "predictive" strategy weights could be obtained, they would themselves form another valuable data sample, well worth separate investigation (consider hypotheses like "influential people tend to pay less than average attention to the others' opinions", etc.)

4. Different statistics have already been obtained for every user. It seems natural to look for correlations between them (consider hypotheses like "the users who post more have higher authority", "influential people have higher probability of obtaining new connections", etc.) Another interesting step would be to classify the users into some groups, e.g. based on their posting habbits (this could be achieved by means of some self-organizing map, such as Kohonen network). Some social phenomena such as homophily [14, 15] suggest clustering effects.

5. Statistics reveal some unusual cases (e.g. some users post a lot of messages, but receive considerably less replies than others). The underlying social mechanics could be understood better, if such cases are studied carefully.

6. The real communication graph was used as an input for our model. At the same time, it could also be simulated using the empirical information about distributions of node degrees and reply times. The algorithms for building such graph could be as simple as "reply to the latest message" (see Appendix Y). An alternative approach for modeling communication graphs would be the use of some preferential attachment algorithm, combined with some kind of "emotional energy conservation law", required for controlling the shape of degree distribution. Such communication graph simulator will significantly increase the model's flexibility, making it possible to variate its size, time scale and other important properties.

7. The model was kept as basic as possible in order to simplify results interpretation. One of the most obvious improvements is use of multiple events memory for the agents (in the current model when one agent replies to another, he "remembers" only their latest communication event, but not the whole history). The concept of "current mood" for each agent (self-explanatory) also seems reasonable, as well as making decisions based on extended set of parameters, e.g. (here A makes a decision of what to reply to B) "what A friends think of B friends", "what A thinks of B friends", or even "what B friends think of B".

8. Known fraction of messages with zero replies can be used as a rough approximation for the user's authority. Of course, such substitute has its own limitations, however this data can be obtained from the communication graph automatically. This valuable feature unlocks whole new range of research opportunities, allowing for "batch" forum processing.

Conclusions

1. Parsed the forum, estimated parsing accuracy. An algorithm for reverse-engineerig the communication sequence for generic phpBB forum was designed, implemented and tested on a real forum. A subjective estimate of its accuracy was obtained – the algorithm scored 77% in correct replies identification, and produced no “false positives”, which makes it usable in scientific research.

2. Introduced some new constructs like communication graph and formalized the concept of authority.

3. Obtained wide range of descriptive statistics. Namely posting volume grouped by user, posting volume grouped by date, degree distribution in relationship graph, distance distribution in relationship graph, reciprocity, number of replies per message, posting rhythm, join and leave times and triad significance profile. Some of those are “standard” methods of analyzing social networks, while others are not as popular in scientific research. As expected, statistics suggest “small world” structure of the relationships between users, with 95% of all users being separated from each other with maximum 3 handshakes. Communication is highly reciprocal and posting volume time series is very volatile. Studying delays between replies allowed to obtain interesting information about “posting habbits” of different users, while analyzing the number of replies received for each message allowed to find unusually “popular” and “unpopular” users. Triad significance profile is rather special, clearly demonstrating that reciprocal relationships are preferred in this network.

4. Designed a simple analytical model and studied its properties. Four simple strategies were suggested (reply the same as before, Tit for Tat, reply the same as community, transitive reply the same as community). Although this model has some severe limitations, it allows to formalize whole communication process as a relatively simple six-dimensional map. This map was iterated for some sample configurations, and its stationary points were found for a special “fully reciprocal” case. The “common enemy” configuration is stable.

5. Created a computational model (simulation) and studied its results. This model is essentially an extension of the previous one, adapted to realistic setting. Real communication graph was used as an input, but emotional payload was generated randomly. The major questions was whether this model behaves predictably or chaotically as we variate user replies slightly. It appears that small changes in replies cause small changes in long-term authority distribution, thus this model is stable in some sense. Furthermore, it seems that the “transitive reply the same as community” strategy can even damp errors in replies to certain extent.

6. Outlined directions for future research. Internet forums are new and interesting target for research. In the last section there are some of the ideas that couldn’t be implemented in scope of this work due to lack of time.

Appendix A: Technical Details

There are seven major software components that were used in this work:

1. Database: Oracle 11g XE (Express Edition) is used as a RDBMS engine. Data schema is trivial – all raw data is stored in the following tables: TOPICS (ID (PK), WHEN, TITLE), INTERACTIONS (TOPICID (PK, FK), SEQ (PK), WHEN, QUOTES, FROMUSER, TOUSER, TEXT), RATINGS (TOPICID (PK, FK), SEQ (PK, FK), WHEN, TYPE, VALUE, RATERID (FK)), RATERS (ID (PK), NAME, PASSWORD). At the same time there is a number of rather complex queries and stored procedures for gathering statistics and performing the simulations, implemented in SQL and PL/SQL respectively. A simple JDBC access layer was chosen as the most adequate solution.

2. Network parser: Implemented in Java using regular expressions, Apache HTTPClient and HTMLUnit libraries. To avoid overloading the server there is a five second delay between HTTP requests. In case of any network problems the requests were resent maximum 10 times and the parsing state was persisted in the database, so that the process could be interupted at any time and then resumed from the place where it stopped.

3. Raw data analyzer: this component is responsible for reconstructing the in-memory representation of the forum, based on the raw communication data read from the database. Data model is implemented in Java as a set of POJOs, generated by JAXB from the XML Schema. This approach simplifies serialization of the objects such as topics and messages, which now can be transferred over the network in XML format (see below).

4. Simulation: the model itself is implemented as a set of basic Java classes. All data is pre-loaded in RAM before the simulation begins, and primitive types (such as int and double) are used wherever possible to increase computational performance.

5. Statistics calculation: although this component represents the major part of this research, its implementation is very straightforward – some SQL queries executed using JDBC layer from Java code, which then persists the results in simple CSV, TSV (tab-separated values) or LST (*Mathematica* list) files. A couple of PL/SQL stored procedures are used for the most complex queries. The queries are designed and tested in Oracle SQL Developer before being wrapped in JDBC.

6. User interface for experiments involving human experts: this part allows the users to “vote” for the messages’ emotional content. It is a simple database-centric web application, written mostly in Javascript using ExtJS framework. On the server side it is backed by some AJAX servlets transferring data in XML format, serialized with JAXB. The application is simple enough to run in a servlet container such as Apache Tomcat.

7. Visualization: all plots in this work were generated by Wolfram *Mathematica* 8.0. A number of notebooks were used to read, process and visualize the data, aggregated by other components. This report is prepared in *Mathematica* as well.

Overal the implementation consists of 162 Kb of Java code encapsulated in 55 classes, as well as 27 *Mathematica* notebooks. It took about 70 hours to parse complete forum, resulting in the database size of about 1.3 Gb, including all indexes. Fetching complete database into RAM takes approximately 10 seconds on Core i7 PC with conventional hard drive and 18 Gb of RAM, consuming about 3 Gb out of it. The slowest operations were:

1. Calculation of TSPs – took about an hour to check all possible triad combinations for sample data, as well as for multiple random graphs;

2. Model experiment – took about five minutes for each simulation run for 2 month period, mostly due to costly authority calculations on each step (for each message). In total the simulation ran for about 120 hours in order to obtain smooth statistics for all required parameter combinations.

It has to be noted that most of the operations are trivially parallel and do not require access to external system resources, thus computation performance scales almost linearly when executed on multiple cores. For example it was possible to run four simulations with different set of parameters in parallel with minor time penalty (it took only about 20% longer, resulting in 320% of overal performance improvement).

The software was implemented using the standard industrial toolset, including Eclipse and Oracle SQL Developer IDEs, Trac issue tracker and Subversion source control system. Wolfram *Mathematica* 8 was used as an exceptionally powerful visualization platform and WYSIWYG editor for this report.

Appendix B: Modeling Communication Graph

Let us take a quick glimpse at one of the simplest approaches towards modeling communication graphs (for more sophisticated algorithms refer to [9]): “reply to last” algorithm, and it essentially means exactly that – every next message in a conversation is a reply to the previous message with some probability P, or a simple stand-alone message with probability $(1 - P)$. P is taken to be the same as the fraction of replies in real-world forum (about 43.7% in our case), and every next message is written by the random user according to the same probability distribution as observed in the real-world forum. This means the number of users, their posting frequencies (basically the shape of the graph on Figure 6) and fraction of replies are taken as inputs to the model, and thus remain the same as in the real forum.

After graph is generated, its statistics can be compared to the real ones, calculated before.

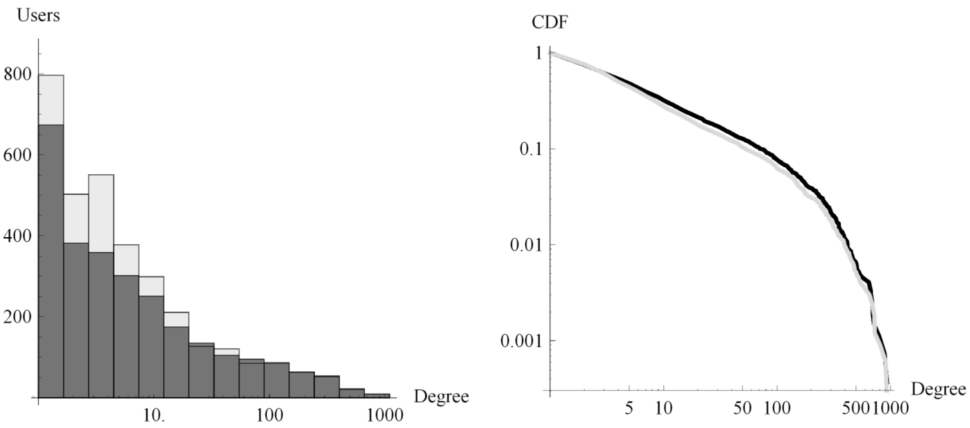


Figure 27. Degree distribution in the relationship graph (analogue of the Figure 7) for communication sequence modeled using “reply to last” approach. We can see that although the graphs look similar to the originals, the spike representing large number of users with low out-degree is missing.

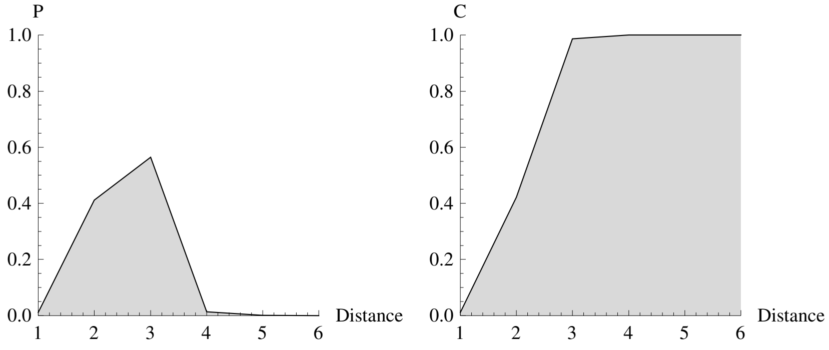


Figure 28. Distance distribution in the relationship graph (analogue of the Figure 9) for communication sequence modeled using “reply to last” approach. Number of “four handshakes” links is lower, while number of “two handshakes” is higher for the simulated graph, compared to the original.

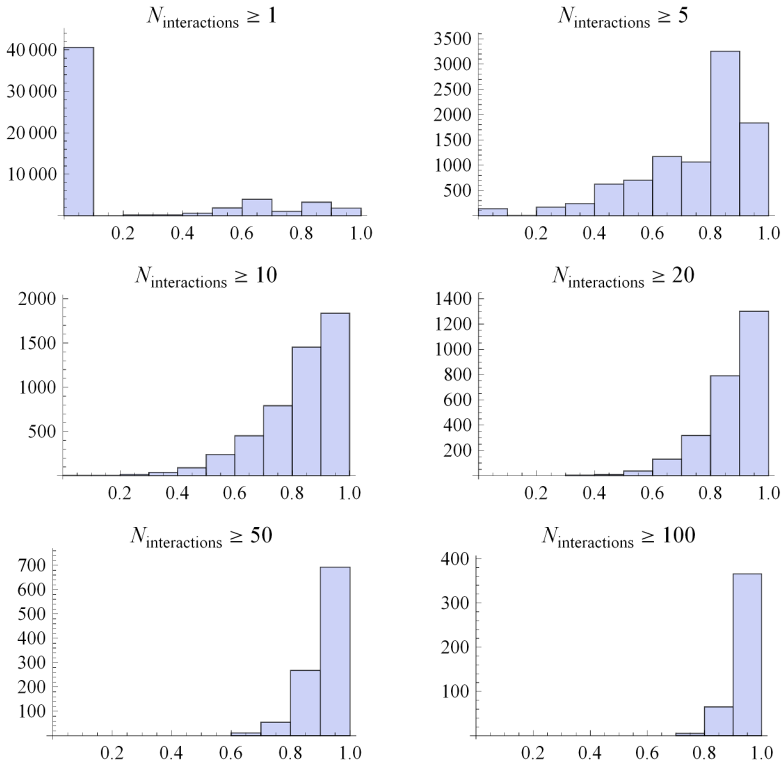


Figure 29. Reciprocity histogram (analogue of the Figure 10). The spikes are more evident, i.e. communication is more reciprocal (which is not surprising at all, taking into account the way we modeled the sequence).

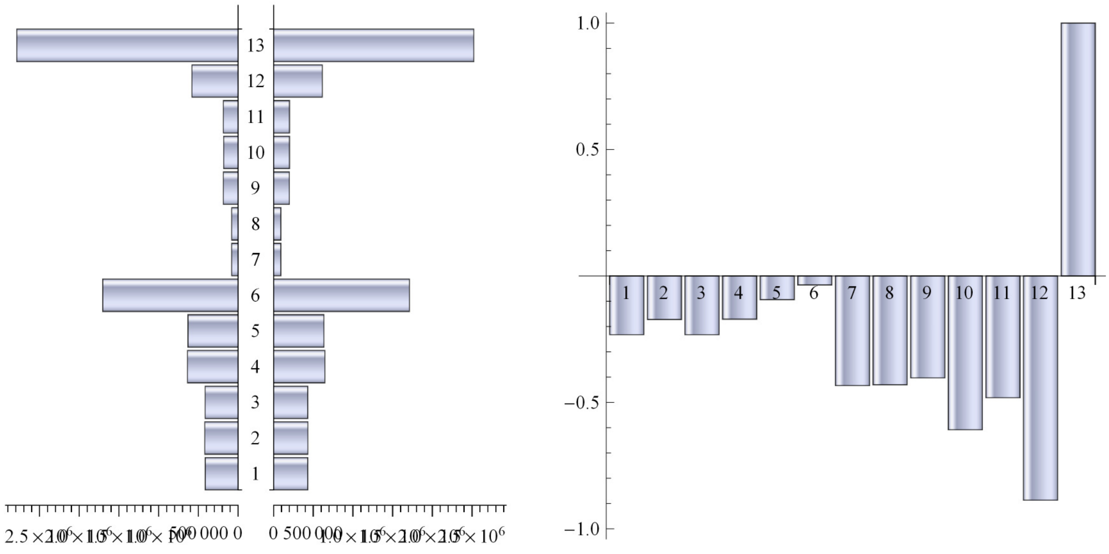


Figure 30. Triad significance profile. Remember that the data for the right plot is normalized. In this case it is easy to see that it makes little sense, because the left plot clearly suggests the triad frequencies are the same as for the random ensemble with the same degree distribution.

As we can see from the graphs above, the simplest "reply to last" algorithm delivers a reasonable approximation for the relationship graph (most of the differences in statistics have quantitative character). However triad significance profile suggests the major qualitative difference, making this algorithm barely useful for modeling relationships.

References

[1] H. Jeong, Z. Neda, and A.-L. Barabasi. *Measuring preferential attachment for evolving networks.* Europhysics Letters, 61, 2003.

[2] M. Peltomaki, M. Alava. *Correlations in bipartite collaboration networks.* Journal of Statistical Mechanics, P01010, 2006.

[3] G. Kossinets, D. J. Watts. *Empirical Analysis of an Evolving Social Network.* Science, 311:88–90, 2006.

[4] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, B. Bhattacharjee. *Growth of the Flickr Social Network.* WOSN'08, Seattle, Washington, USA, August 18, 2008.

[5] J. A. Bargh, K. Y. A. McKenna. Ann. Rev. Psych. 55, 573, 2004.

[6] B. M. Gross. Paper presented at the First Conference on E-mail and Anti-Spam (CEAS). Mountain View, CA, July 30-31, 2004.

[7] B. Freiesleben de Blasio, Å. Svensson, F. Liljeros. *Preferential attachment in sexual networks.* PNAS June 26, 2007, vol. 104 no. 26, pp. 10762–10767.

[8] R. Merdoc, 1968. Science 159:56–63.

[9] J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins. *Microscopic Evolution of Social Networks.* KDD'08, Las Vegas, Nevada, USA, August 24–27, 2008.

[10] R. Milo et al., Science 298, 824, 2002.

[11] R. Milo, et al. *Superfamilies of Evolved and Designed Networks.* Science 303, 1538, 2004.

[12] P. Holland, S. Leinhardt, D. Heise, Eds., in Sociological Methodology (Jossey-Bass, San Francisco, 1975), pp. 1–45.

[13] I. Bezáková, A. Kalai, and R. Santhanam. *Graph model selection using maximum likelihood.* In 23rd ICML, pages 105–112, 2006.

[14] S. Jones. *The Internet Goes to College: How students are living in the future with today's technology.* Pew Internet & American Life Project, 2002.

[15] N. K. Baym, Y. B. Zhang, M. Lin. New Media & Society 6, 299, 2004.