ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Lecture with Computer Exercises:
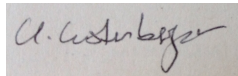# Modelling and Simulating Social Systems with MATLAB

Project Report

Simulation of Information Spreading in a Facebook Network

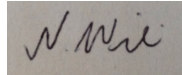Urs Lustenberger , Nino Wili , Patrick Zöchbauer

Zurich
December 2013
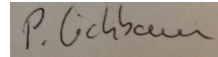
Agreement for free-download

We hereby agree to make our source code for this project freely available for download from the web pages of the SOMS chair. Furthermore, we assure that all source code is written by ourselves and is not violating any copyright restrictions.

Urs Lustenberger   Nino Wili   Patrick Zöchbauer

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of Originality

This sheet must be signed and enclosed with every piece of written work submitted at ETH.

I hereby declare that the written work I have submitted entitled

Simulation of Information Spreading in a Facebook Network

is original work which I alone have authored and which is written in my own words.*

Author(s)

Last name	First name
Lustenberger	Urs
Wili	Nino
Zöchbauer	Patrick

Supervising lecturer

Last name	First name
Kuhn	Tobias
Woolley	Olivia

With the signature I declare that I have been informed regarding normal academic citation rules and that I have read and understood the information on 'Citation etiquette' (http://www.ethz.ch/students/exams/plagiarism_s_en.pdf). The citation conventions usual to the discipline in question here have been respected.

The above written work may be tested electronically for plagiarism.

Zurich, December 13, 2013
Place and date

see below
Signature

Print form

3

Contents

1   Abstract

A model for the information spreading in a social network was developed and investigated. The SIR model known from epidemiology was adapted to this problem. A homogeneous system, where the evolution is described by differential equations, was compared with an agent-based simulation. It was found that there are cases where they significantly differ. In the agent-based simulation, individuals that are way more important than the rest were identified. It was found that the betweenness centrality might be a key parameter to determine such influential individuals.

2   Individual contributions

Nino played the key role in developing our simulation. Patrick was in charge of the data-fetching. Urs was responsible for the visualizations. The analysis was preformed together.

3   Introduction and Motivations

3.1   Introduction and fundamental questions

Social networks like facebook and twitter are growing fast. Most young people have an account in at least one of them. As many articles, pictures and videos on the internet can be shared easily, it is a very interesting question how this information spreads in such networks. First, the internet may have become the fastest and sometimes most important source of "news" and information besides "traditional" media and second, companies have an increasing interest in targeting the right people with their adverts.

The simplest way to describe the information spreading uses similarities to epidemiology, which means that the flow of information of one person to another is looked at as an "infection". In this work, we implemented a simple model in a homogeneous way with a set of differential equations and their numerical solutions as well as an agent-based, heterogeneous model, using a real facebook network.
The fundamental questions of this work are:

- Are there relevant differences in the time evolution of the homogeneous and the agent-based model?

- Are there indiviuals which are more important to the spreading of information (also called influentials)? Can they be recognized in the sense of position and

connectivity in the network?

3.2 Motivation

As the team of authors consists of two chemists and a mathematician, the personal motivations also have a broad range. We were glad to recognize parallels between our simulations with statistical physics , chemical reaction kinetics, and graph theory.

4 Description of the Model

The process of information spreading has many similarities with epidemiology, therefore the well known SIR-model was adapted[1]. The "susceptibles", are not aware of the information and are called "ignorants" within this work. "Infected" individuals know about the information and are willing to share it with other people, in other words they spread it and are therefore called "spreaders". The adaptation of the epidemiological term "recovered" is not that straightforward as its meaning in this context is not entirely clear. However, they can be interpreted as individuals being aware of the information but do not share it with others. They are called "stiflers".

In the SIR-model as well as in the agent-based model, the following set of "reaction equations" was used (I: Ignorant, S: Spreader, R: Stifler):

$$
\begin{cases}
I + S \xrightarrow{\lambda} 2S & (1) \\
S + R \xrightarrow{\alpha} 2R & (2) \\
S + S \xrightarrow{\alpha} S + R & (3)
\end{cases}
$$

The main difference to the standard SIR-model is that spreaders do not become stiflers spontaneously. This change is only induced by meeting another spreader or stifler.

4.1 Homogeneous SIR model

In the mathematical treatment of the homogeneous model, the set of differential equations is expressed in terms of the densities $i(t) = I(t)/N$, $s(t) = S(t)/N$ and $r(t) = R(t)/N$, where N is the number of individuals in the network.

$$\begin{cases} \dfrac{\mathrm{d}i(t)}{\mathrm{d}t} = -\lambda \cdot s(t)i(t) & (4) \\[2mm] \dfrac{\mathrm{d}s(t)}{\mathrm{d}t} = \lambda \cdot s(t)i(t) - \alpha \cdot s(t)[s(t) + r(t)] & (5) \\[2mm] \dfrac{\mathrm{d}r(t)}{\mathrm{d}t} = \alpha \cdot s(t)[s(t) + r(t)] & (6) \end{cases}$$

4.2 Agent-based model

In the inhomogeneous, agent-based model the individuals are connected in a certain manner (facebook friends in this particular work). Only connected agents are able to meet. If a meeting occurs, transitions are induced at a certain probability depending on the relation between the agents (details are discussed later). Additional to the different connectivity of the agents, they also have a different activity, i.e. different probability to meet somebody.

To convert the reaction equations into an agent-based model, time was discretized and in each time step a series of two steps is performed. First, the agents randomly meet another agent they know or nobody. Only two-agent meetings are possible. In the second step, the status of the agents change corresponding to the situation. There are six possible combinations of ignorants, spreaders and stiflers. Three of them correspond to the "reactions" 1-3, the other ones (I+I, I+R and R+R) have no other influence than "occupying" the agents. The probability λ was chosen to be dependent on the number of common friends. For simplicity, α was a constant.

5 Implementation

5.1 Homogeneous SIR-model

The system of ordinary differential equations described in section 4 was solved in MATLAB using ode45. The simulation was carried out four times using α to λ ratios between 0.1 and 10. The starting conditions were always S = 383, I = 1 and R = 0.

5.2 Agent-based model

Initial condition At the beginning of each simulation, all agents are ignorant but one, which is a spreader. In the biggest performed simulation, every one of the 384 nodes was the first spreader 10 times.

Determine the meetings   In order to determine who meets whom, the program goes through the vector of agents (1:N) randomly. With a probability corresponding to the activity of that agent, he may meet somebody. The person he meets is determined randomly and must be one he knows and one which is not already meeting another agent in this time step. See `talkstep.m` for details.

Status changes in meetings   After the meetings are determined, the status of each agent has to change according to the model. In order to be able to determine who infected whom, and how many were infected, a list `infecpath` was created in each round containing directed edges of infections.

Exit condition   The current simulation round was terminated if further spreading was not possible. This is the case when all current spreaders and all their friends have an activity of zero, including the obvious case that the number of spreaders is zero. For stability, not more than 5000 steps of meeting and status updates were performed in each initiated round.

Running the simulation   3840 rounds were initiated. For every round the first infected person, the number of meetings, the number of direct infections and the number of cumulative infections were saved for further analysis. A separate simulation was carried out to investigate the influence of α.

5.3   Importance of an individual

One of our main questions was, if there are individuals which are more important to the spreading of the information. The definition of importance in this context is not trivial. The nearest quantitative variable that might indicate the importance is the number of *direct infections*, meaning the sum of all meetings of an infected person which resulted in the infection of his meeting partner. But in our opinion, a better variable is *cumulative infections*. This is defined as the number of direct infections of a person plus the number of direct infections of all other agents infected by that person and so on (the whole subtree). Watts [2] defines a similar value, namely the size of a *cascade*. A cascade is the sequence of infections initiated by an individual. The size of a cascade is the number of infections in this cascade. So the value of cumulative infections and the size of the cascade are the same if only considering the person which was infected first (our initial condition). Cumulative infections are in a sense more general.

Determine cumulative infections With a given `infecpath`, containing the directed edges of infections, it is easy to obtain the number of direct infections of every node. It is also straight-forward to get the number of cumulative infections. The ladder was obtained in a recursive manner in the following way:

```
1     L=length(infectpath(:,1));
2      for  i=L:−1:1
3
4         p1=infectpath(i,1);
5         p2=infectpath(i,2);
6         cum_infections(p1)=1+cum_infections(p1)+cum_infections(p2);
7      end
```

5.4   Get the network

To get the data for our simulation, we used the so-called Facebook Graph API, a programming tool designed to support better access to conventions on the facebook social media platform.[1] Unfortunately, there is no implementation for MATLAB, so the data-fechting was programmed in python using facebook sdk.[2]

Get access First, one needs to get access to the social graph. Therefore, an access token is required.[3]

Get graph Having enterd a valid token, one can access all information this token provides. For simplicty, only the non-anonymized case is discussed in detail (for more details on the anonymization, look at subsection anonymization).

```
1  profile = graph.get_object("me")
2  friends = graph.get_connections("me", "friends")
```

For this work not only data of ones own profile is of interest, but also all available information one gets from his friends. The object friends is an array containing all id's and data of the friends. Since the focus of this work does not lie primary on the direct friends, but the connection in between them, we have to access at least the mutual friends.

---

[1]http://www.techopedia.com/definition/28984/facebook-graph-api
[2]https://github.com/pythonforfacebook/facebook-sdk
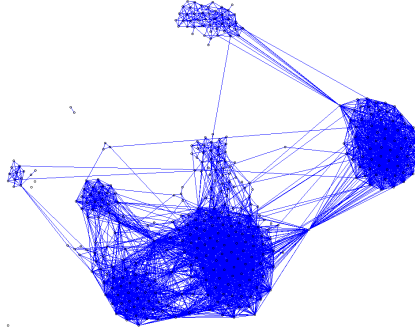[3]https://developers.facebook.com/docs/facebook-login/access-tokens/

Figure 1: A graph of the used facebook network with all edges.

```
1   mutualfriends = graph.get_connections(tfriend, "mutualfriends")
```

Get activity   For each friend the number of posts per day is determined. This ratio indicates the activity of each individual on facebook.

Get coordinates   To visualise the network in MATLAB, the software "Gephi" was used to determine the x- and y- coordinates of each individual. Therefore, the file "gephi.csv", an edge list of the network, was imported. The nodes were then arranged in communities using the layout method "ForceAtlas2". Finally, normalised node coordinates where exported to "coordinates.gdf". After deleting all headers and all edge data from this file, the coordinates could be easily imported to MatLab. The Network is visualized in Figure 1.

Anonymisation   The anonymisation is acutally only a pseudo-anonymisation. First all id's get arranged in order of size. Secondly, a list from 1 to n (number of friends) is created. One then defines a bijection between the two sets, such that the i-th element of the first set, maps to the i-th element of the second set. Even though, this is not really anonymising our data. It is more then sufficient for our purpose, the privacy.

6 Simulation Results and Discussion

6.1 Characteristics of the network

The used network consists of 384 nodes. The distribution of the number of friends and the local clustering coefficients are shown in figure 2. The one of the betweenness centrality in figure 3.
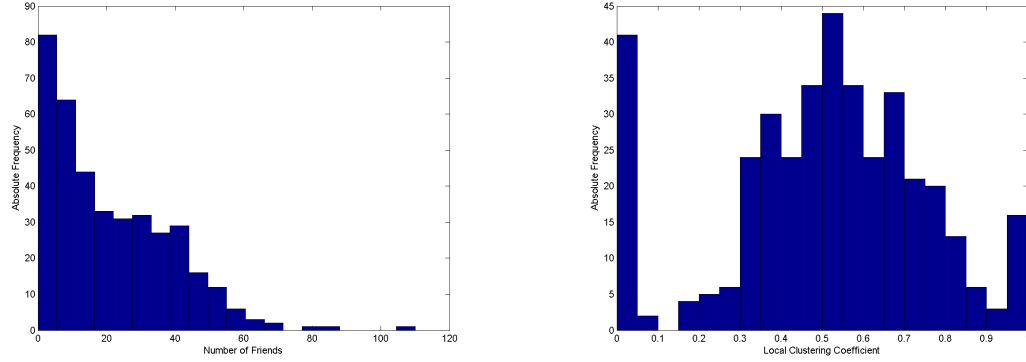

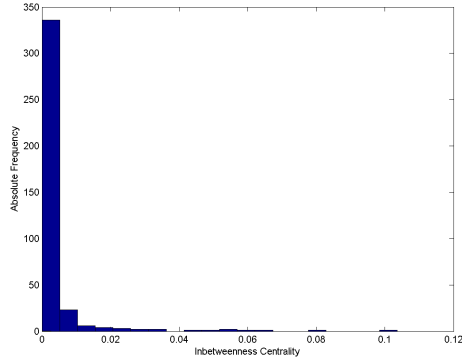Figure 2: Distribution of the number of friends (left) and the local clustering coefficient (right).

Figure 3: Distribution of betweenness centrality.

The main problem of our network is, that it consists out of all the friends of one of the authors, with the author removed. The probability of infecting another person is dependent on the mutual friends. Thus, it is possible that two individuals might have many common friends in reality, but the contrary appears because the author does not know these common friends.
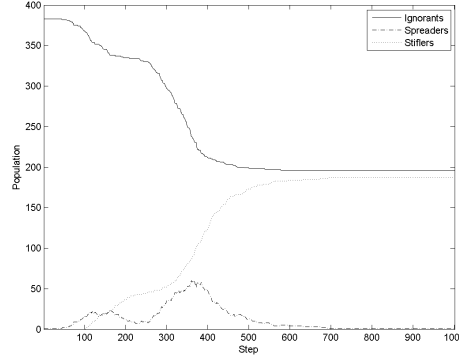
Figure 5: In the agent based model several local maxima of function S(t) can occur.

6.2 Differences in time evolution

Most of the time, the form of the population profiles of the homogeneous and the inhomogeneous, agent-based model are very similar (Figure 4).
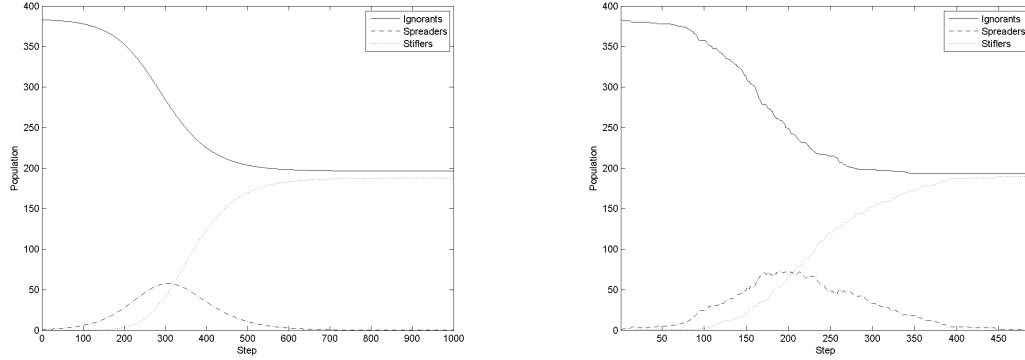

Figure 4: Left: Typical evolution of the system in the homogeneous SIR-model. Right: The evolution in the agent-based model. In most cases, they coincide.

There were also interesting cases where two or more local maxima occurred, as in figure 5. This can never occur in a homogeneous model. This requires some special conditions e.g. loosely connected neighbourhoods. This might be a rather obvious insight, but it proofs that a homogeneous model might fail representing reality.
In the homogeneous model the function S(t) (number of spreaders) can only have one local maxima due to mathematical reasons.

6.3 Influence of α

In the big simulation conducted α was assigned the value 0.3. This value was chosen because higher values limit the information spreading to a large number of individuals, whereas small numbers increase computing time significantly. In another simulation the influence of α was investigated by simulating the information spreading 25 times for 20 different values of α between 0.05 and 1, while starting always with the same spreader. In figure 6 the average number of remaining spreaders at the end of the simulation is plotted against α. Figure 7 shows the number of remaining spreaders vs. α obtained with the homogeneous model. The resemblance of those two figures implies that both models behave similarly with respect to α. Even though it was not investigated, we assume that α does not influence the importance of individuals in the network.
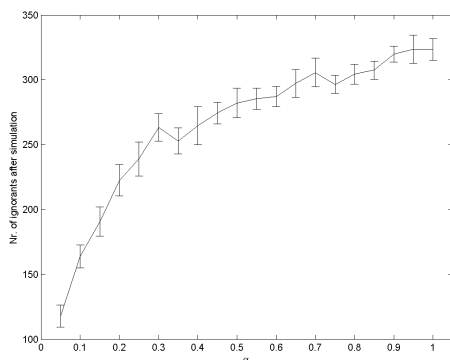

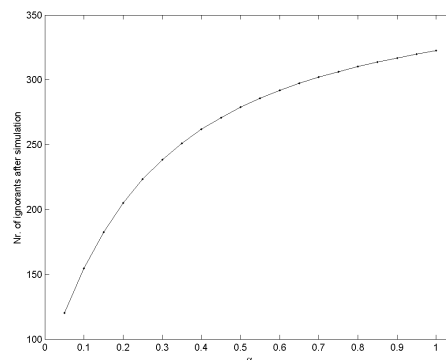Figure 6: Influence of α in the agent-based model.

Figure 7: Influence of α in the homogeneous model.

6.4 Influentials

6.4.1 Existance and Importance of Influentials

Figure 8 is a simple histogram showing that the vast majority of people has only a small cumulative infection value. However, there are some individuals with a very high value. This an indicator for the existence of influentials.

One can even go one step further and try to estimate the importance of influentials on all infections that occurred in total. In order to do so, the individuals are sorted according to their value of cumulative infections. Then, for each person m, we sum up all the infections that occurred in rounds where m itself did not infect anybody.
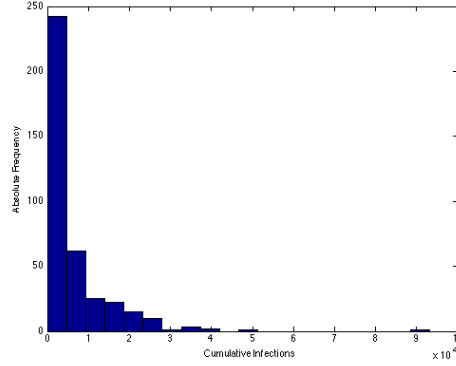
Figure 8: Distribution of the cumulative infections, summed over all 3840 rounds.

This function is called f(m). Figure 9 shows the value of

$$g(m) := \sum_{i=1}^{m} f(i)$$

This can be interpreted as the number of infections that occurred in rounds where all persons more or equally important than m did not infect anybody. So $g(1) = 0$ and $g_{max} =$ total infections.

Given these results, we get that 94 % of all infections happen in the rounds where the 1 %-quantile of the most important people were not excluded.

This result seems to proof the existence of influentials in our network. However, one could say that this only shows that the so-called influentials are just contributing in rounds where many persons get infected, not that this is caused by the influentials. But they are also the ones with the most cumulative infections, which shows that they really *cause* a high number of infections.

6.4.2   Determine influentials

The next question is, whether it is possible to determine parameters indicating influentials. e.g. the number of friends or the cluster coefficient.

We were not able to identify a single parameter clearly predicting the influentials. However, we recognized that individuals with a higher number of friends and a high activity infect more people directly. In addition, they might have a higher number of cumulative infections, but the influentials or the individuals with the highest number of cumulative infections are not strictly following this trend. The best indicator we
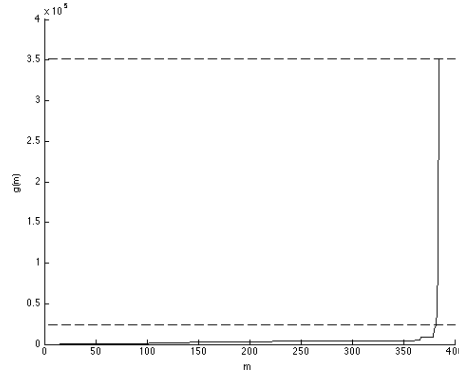
14

Figure 9: g(m) (see above for details). The bottom dashed line indicated the number of all infections in rounds where the four most important individuals did not infect anybody(6% of all infections). The top dashed line shows the total number of infections in all 3840 rounds in the simulation.

found is the *betweenness centrality*: The two individuals with the highest number of cumulative infections are the two with the highest betweenness centrality (figure 10).
But a higher betweenness centrality does not indicate a higher importance. Individuals with a higher betweenness centrality connect different neighbourhoods, so their importance may come from the fact that they are one of the few individuals which are able to connect large groups of persons which would be isolated without these linkers.
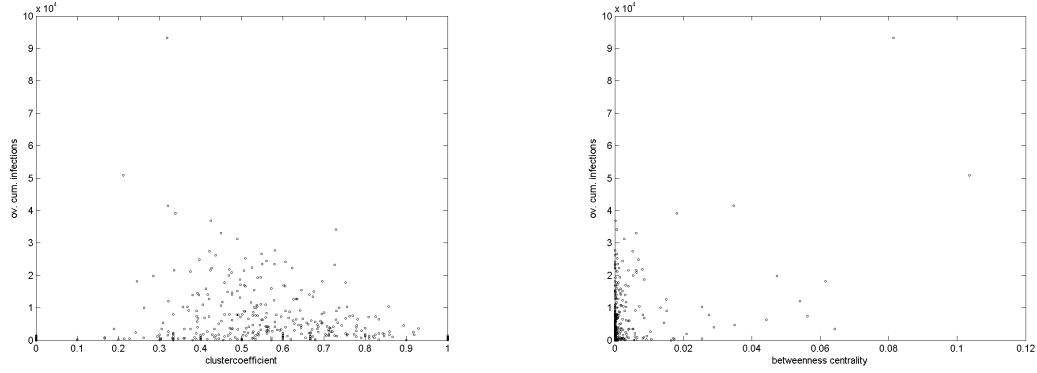
15

Figure 10: Left: Cumulative infections in all rounds vs. the cluster coefficient. A higher cluster coefficient does not indicate a higher number of cumulative infections. On the contrary, within our model, very clustered persons can become spreaders fast, but they also become stiflers fast, limiting the possibility that the "inforamtion" reaches other neighbourhoods. Right: Cumulative infections vs. betweenness centrality. The two persons with the highest betweenness centrality also have the highest number infections.

7   Summary and Outlook

- Under certain circumstances the population profiles of the agent-based model differs significantly from those of the homogeneous model.

- Individuals being more important for information spreading, in terms of cumulative infections, were found. The betweenness centrality might be used to roughly predict those especially influential individuals.

Future work on this topic, should focus on getting more advanced data. Such that, the problems mentioned in section 6.1 can be avoided. Obviously, it would also be of interest to preform the experiment with more individuals. e.g. several connected networks. Further analysis can be done in the dependence of the parameters e.g. λ on characteristics of the network. Also the parameter α could be investigated in more detail. Further on, to compare or even confirm the results, it would be desirable to gather real world data.

8 Appendix

References

[1] A. Barrat, M. Barthlemy, A. Vespignani. Dynamical Processes on Complex Networks. *Cambridge University Press* Chapter 10, pp. 216-241

[2] D.J.Watts, P.S.Dodds. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research* Vol 34, No. 4 (December 2007), pp. 441-458

[3] J.Goldenberg, B.Libai, Eitan Muller. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters* 12:3, 211-223, 2001
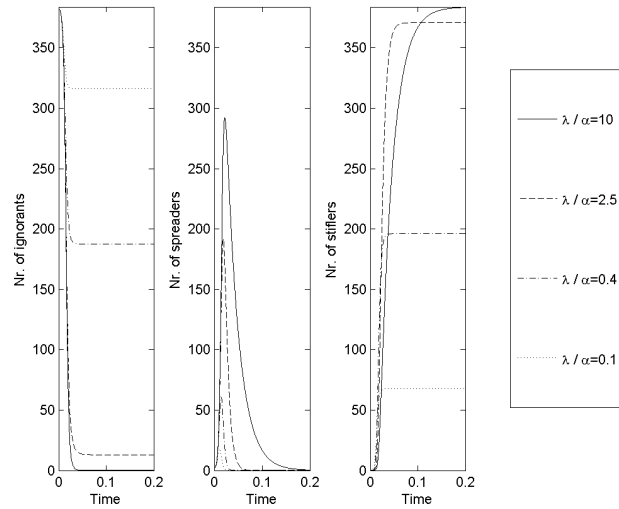
8.1 Additional figures

Figure 11: Evolution in the homogeneous SIR-model for different λ/α ratios.
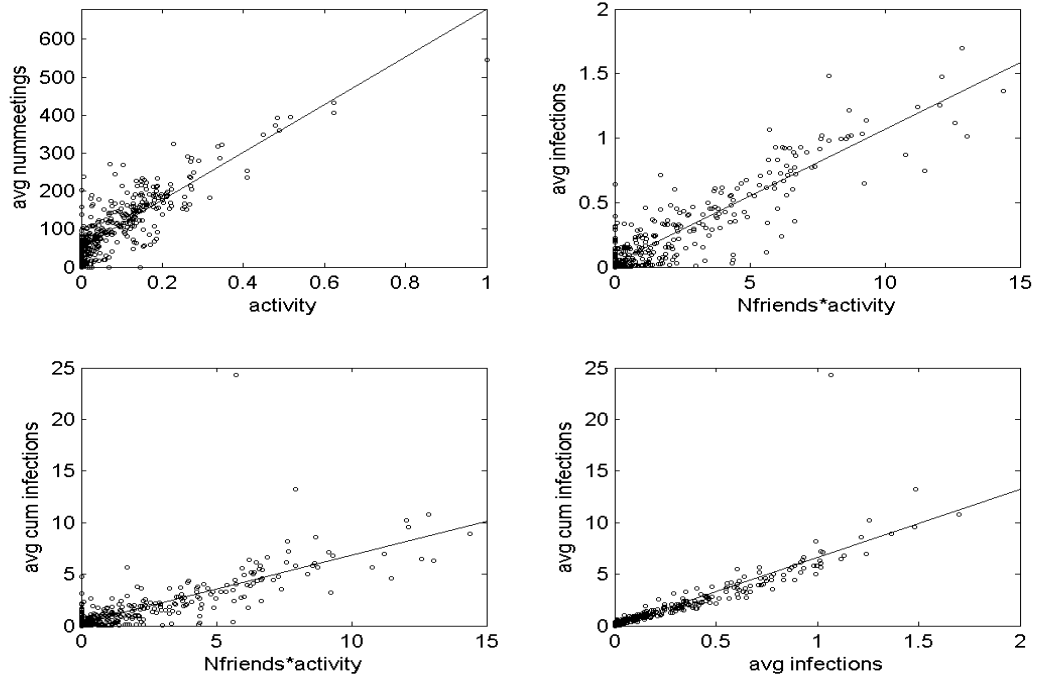
17

Figure 12: Top left: Average number of meetings vs. activity. The trend line's slope is $630 \frac{Avg.\ nummeetings}{activity}$, $R^2 = 0.71$. Top right: Average number of infections vs. number of friends times activity. The trend line's slope is $0.10 \frac{avg.infections}{Nfriends*activity}$, $R^2 = 0.82$. Bottom left: Average number of cumulative infections vs. number of friends times activity. The trend line's slope is $0.66 \frac{Avg.\ cum.infections}{Nfriends*activity}$, $R^2 = 0.62$. Bottom right: Average number of cumulative infections vs. number infections. The trend line's slope is $6.6 \frac{avg.\ cum.\ infections}{avg.\ infections}$, $R^2 = 0.82$.