Probing models of information spreading in social networks

J Zoller and S Montangero

Institut für Quanteninformationsverarbeitung & IQST, Universität Ulm, 89069 Ulm, Germany

E-mail: simone.montangero@uni-ulm.de

Received 17 March 2014, revised 9 September 2014
Accepted for publication 12 September 2014
Published 9 October 2014

Abstract
We apply signal processing analysis to the information spreading in a scale-free network. To reproduce typical behaviours obtained from the analysis of information spreading in the World Wide Web, we use a modified SIS (from ‘susceptible–infectious–susceptible’) model where synergy effects and influential nodes are taken into account. This model depends on a single free parameter that characterizes the memory time of the spreading process. We show that by means of fractal analysis it is possible—from aggregated easily accessible data—to gain information on the memory time of the underlying mechanism driving the information spreading process.

Keywords: complex networks, fractal analysis, SIS model
PACS numbers: 64.60.aq, 89.65.-s, 95.75.Wx

(Some figures may appear in colour only in the online journal)
The efforts made in the aforementioned analysis are enormous, as usually a huge amount of data have to be processed and extensive numerical simulations on large network sizes have to be performed to support and verify the theoretical analysis. Indeed, this is a typical problem encountered in complex system analysis, where microscopic models of the actions of the elementary constituents or agents (which may be brain cells, ants, or market agents) are rarely known in detail; nor are simulations of the whole system evolution based on first principles possible: for example, simulating brain activity, including a detailed description of each single neuron, has been considered—until very recently—practically impossible [19]. A possible approach for overcoming these limitations adopted in standard complex system analysis is studying time correlations of signals extracted from a complex system or part of it—e.g. brain, social or stock market activities—and inferring important information on the global status of the system or on the ongoing processes [20]. Working along these lines, one of us introduced a tool for investigating the properties of time series extracted from the evolution of social networks and applied it in analysis of the World Wide Web [21]. In particular, the method introduced in [21] is based on the working hypothesis that the correlations present in time data series are representative of the activities of the underlying communities, and thus from studies of them it is possible to indirectly infer properties of the agents themselves and of the interactions between them. It has been shown that correlations can be quantified by means of the fractal analysis of the signal [28], and argued that a fractal signal corresponds to a strong active community, possibly very influential and with high probability of lasting for a long period of time.

In [21] this approach has been introduced heuristically and applied in the analysis of real-world time series, i.e. occurrences of keywords in the World Wide Web; in this work we follow a bottom-up approach in order to test a working hypothesis and to look for theoretical models of the fundamental mechanism responsible for the creation of such correlations compatible with the experimental data. We apply the fractal analysis to different models of information spreading in scale-free networks and we show that it is possible to discriminate between them. In particular we show that the standard SIS model [2] is an oversimplified model for that purpose, as it does not reproduce the real-world scenario and thus has to be rejected. The SIS model is very well suited for describing the spreading of viruses and illnesses; however—as we will show in the following—the SIS model does not reproduce the rich behaviour observed in World Wide Web data. Indeed, as we focus on information spreading in social dynamics, the piece of information is a rumour, useful knowledge, a marketing announcement or a political or philosophical idea; the infected nodes are spreaders that try to convince their neighbours to adopt or follow their suggestions. It is then natural to include in the model the fact that many spreaders will most likely be more effective in spreading a rumour than a single one (synergy) and that spreaders might have different rates of success (influential nodes): an opinion maker idea is more likely to be followed by the whole community, the market leader’s new campaign will spread easily and more successfully than that of an unknown brand (normalized with respect to the investment), and news launched by an important medium will most likely be reported by other media. We introduce a modified SIS model, where synergy components in information spreading [14] and the presence of influential nodes [22] are naturally taken into account: we show that the fractal properties of the data series depend only slightly on the fine details of the network, while they depend strongly on the recovery time of the network nodes. Finally, we show that the modified SIS model is compatible with real data and thus it might allow us to infer important properties for the information spreading process, e.g. the recovery time of the nodes.
1. The model

Throughout this paper we will consider two different models of information spreading. One is the standard SIS model [2], where the nodes have two possible states: infected and susceptible to infection. At every time step, every node that has an infected neighbour has a probability $\nu$ of becoming infected. After a typical time $\mu = 1/\delta$ (number of steps), where $\delta$ is the probability of recovering, the node recovers and become susceptible again. In the modified SIS model that we introduce here, we identify the influential spreaders as those that are highly connected, and we set the probability of infecting as proportional to the number of links that a node has: at every step, each infected node will try to infect every neighbour with probability $\nu = z/z_{\text{max}}$, where $z$ is its coordination number (number of links to other nodes) and $z_{\text{max}}$ is the highest coordination number encountered in the network. Moreover, every spreader attempts to infect all of its neighbours every time step for a typical total time $\mu = 1/\delta$, after which it recovers and becomes susceptible again; that is, every node will experience a synergistic effect tending to make it infected proportional to the number of infected nodes between it and its neighbours. Notice that influential nodes are here influential ‘locally’, as witnessed by the high coordination number, and this definition has in principle nothing to do with that introduced in [22] and based on the $k$-shell decomposition (we will refer to influential nodes defined in the latter way as ‘global’ influential nodes). Finally, notice that synergy and influential nodes effects are included as a (local and) linear function of the number of neighbouring infected nodes and the coordination number of the spreaders; and that while the network is not directed (infection can spread in both directions), the probability of infection is asymmetric, reflecting a natural scenario in social dynamics (a symmetric model has recently been studied in [23]).

We consider, independently from the spreading model that drives the dynamics, scale-free networks of sizes $N$ built via preferred attachment with a power law distribution of coordination numbers, $P(z) \sim z^{-2.33}$ [2]. The nodes are ordered by means of the $k$-shell decomposition as follows [22, 24]: the nodes that can be disconnected from the network by cutting a single link belong to the first shell (i.e. only for this shell do they correspond to those with a single link). After eliminating the nodes belonging to the first shell from the network, the nodes that have a single link remaining belong to the second shell. The procedure is repeated until all nodes are eliminated and the last shell is defined. Notice that from the
second shell on, the coordination number $z$ of a node and its shell usually do not coincide. The infection is injected at a random node (either in the whole network or in a predetermined shell) and we monitor the number of infected nodes $N_I$ as a function of the (discrete) time $t_i = i \Delta t$, where $\Delta t$ is a typical timescale of the system. Notice that in this model we have only one free parameter, namely the recovery time $\mu$ of each infected node. A typical result of this dynamics is shown in figure 1, where the status of the network is depicted at a given time, and infected and susceptible nodes are marked.

2. Results

We first focus on the evolution of the infection under the modified SIS model. Typical results are shown in figure 2 (top), where the percentage of infected network $I = N_I/N$ as a function of time is reported for the two different scenarios that we have found: it either dies out very quickly or it becomes permanent and stabilizes around a nonzero level. The latter behaviour resembles what is found also in other models; see e.g. [25]. This is consistent with the general opinion that ‘rumours are hard to kill’ and with the persistence of chain letters. We report in the bottom panel of figure 2 the probability $P_N$ of having a nondying process as a function of the recovery time $\mu$, with starting points in different $k$-shells. As can be clearly seen, infections starting in the inner shells have a higher probability of persisting in the network than those starting in the outer shells, and in general the longer the recovery time, the higher the probability of a permanent infection. Notice also that although in the network there is always a node with 100% success spreading rate, i.e. $\nu = 1$, this does not imply that the infection becomes permanent. That is, a single very connected (and locally influential) node cannot deterministically influence the whole network. Assuming that the infection has not started from a poor spreader, we see however that the critical recovery probability $\delta_c$, above which the infections will not be sustainable, scales as $1 - \delta_c \propto 1/N$ with the system size (data not shown). This behaviour is typical for small-world networks [26].

From now on, we concentrate on the cases where permanent infection occurs, and in particular on the average level of infection $\bar{I}$ and on the time fluctuations $\Delta I = \bar{I} - \bar{I}$, where $\bar{I}$ stands for a time average. We report in figure 3 the infection level $\bar{I}$ as a function of the inverse recovery time $1/\delta$ obtained using the model introduced here and the SIS model (as a function of $\lambda = \nu/\delta$). While the SIS model displays the typical S-shape dependence [27], the modified SIS model results in a strictly concave dependence of $\bar{I}$ on the recovery time $\mu$ until saturation effects come into play. As can be seen in figure 3, this dependence is compatible with a power law scaling (full lines) that can be estimated with the following simple
theoretical arguments: to sustain a permanent infection level, each node has on average to infect another node during the recovery time $\mu$; that is, the probability of one infection has to be $p \geq 0.51$. The probability of infecting one of the $z_i$ neighbours at every step is $\nu_j = z_j/z_{\text{max}}$ and from elementary probability theory we obtain that the overall probability of having at least one successful infection in $\mu$ steps is

$$
\nu^\mu = 1 - \left(1 - \frac{z_j}{z_{\text{max}}}\right)^\mu.
$$

(1)

Given that for the majority of nodes $z_i \ll z_{\text{max}}$, we can expand $p_1$ in a Taylor series, and the condition of having a permanent infection ($p_1 \geq 0.5$) is satisfied by the nodes that fulfil

$$
z^* \geq \sqrt{\frac{z_{\text{max}}}{2\mu}}.
$$

(2)

Thus, the number of nodes that can sustain the permanent infection are those whose coordination number fulfils equation (2) (notice that this condition is completely independent from the $k$-shell analysis). Their number, that is the infection level $I$, can be estimated from the static network properties, as the number of nodes with $z > z^*$ is given by

$$
I = \int_{z^*}^{\infty} P(z)dz = C (\frac{z_{\text{max}}}{z^*})^{\gamma-1};
$$

(3)

here $C$ is a constant, and $P(z) = z^{-\gamma}$ is the distribution of links of nodes in the network; and in our simulations $\gamma = 2.33$. From equations (2) and (3) we obtain

$$
I \sim C \left(\frac{z_{\text{max}}}{2}\right)^{-\eta} \cdot \mu^{\theta},
$$

(4)
where
\[ \eta = (\gamma - 1)/2. \] (5)

By means of numerical analysis for different networks and network sizes, we fit the values of the constants as \( z_{\text{max}} \sim 2.8\sqrt{N} \) while \( C \sim 2.6 \) (data not shown). The infection rate converges to our theoretical prediction for \( \frac{1}{\delta} \geq 10 \), as reported in figure 3 (inset). The exponent in the power law from equation (4) is not obtained by a fit, but is deduced from a very basic property of scale-free networks: their power law parameter \( \gamma \) (see the introduction). Notice that equation (4) is invertible; thus given a network size and an infection level, it allows us to extract the recovery time of the model.

We finally concentrate on the fluctuations of the infection level around the average value \( \bar{I} \); in particular, we analyse the fluctuations by means of the fractal analysis as introduced in [21]. The fractal dimension of a signal can be extracted by means of the modified box-counting algorithm [28]: in the standard box-counting algorithm the fractal dimension \( D \) of the signal is obtained by covering the data with a grid of square boxes of size \( L \). The number \( M(L) \) of boxes needed to cover the curve is recorded as a function of the box size \( L \). The (fractal) dimension \( D \) of the curve is then defined as

\[ D = -\lim_{L \to 0} \log \frac{L}{M(L)}. \] (6)

The modified algorithm follows the same lines but uses rectangular boxes of size \( L \times \Delta \) (\( \Delta \) is the largest excursion of the curve in the region \( L \)). Then, the number \( M(L) = \sum \frac{\Delta}{L} \) is computed. Such a procedure is illustrated in figure 4, where a typical signal is processed. For any curve, a region of box lengths \( L_{\text{min}} < L < L_{\text{max}} \) exists where \( M \propto L^{-D} \). Outside this region, either \( D = 1 \) or \( D = 2 \) is found: the first equality \( (D = 1) \) holds for \( L < L_{\text{min}} \) and it is due to the coarse graining artificially introduced by any discrete time series; the second one \( (D = 2) \) is obtained for \( L > L_{\text{max}} \) and it is due to the finite length of the time series analysed. The boundaries \( L_{\text{min}}, L_{\text{max}} \) have to be chosen properly for any time series, and a power law fit allows us to extract the fractal dimension \( D \). The upper panels of figure 5 show typical results from this procedure. The fractal dimensions measure the degrees of correlations in a time series, as shown for example in [29]: in the case of a stationary Gaussian random process, one can show that if the correlations in the time series are such that \( C(h) = 1 - |h|^\beta \) when \( h \to 0 \) for some \( \beta \in (0, 2] \), then the fractal dimension is related to the exponent as

![Figure 4. A typical time evolution of the number of susceptibles, \( I \). Fluctuations around their average \( \bar{I} \) are analysed by grouping them into rectangles of equal width \( L = t_i - t_j \).](image-url)
Thus, the faster the decay of the correlation, the lower the fractal dimension; for example, $D = 2$ corresponds to the case of very slowly decaying correlations, $\beta = 0$, while $D = 1$ corresponds to quickly decaying correlations, $\beta = 2$.

We performed the aforementioned analysis for a wide range of different evolutions, for the SIS and modified SIS models, for different recovery times $\mu$. The results are presented in figure 5 (lower panels), where we show that the SIS model gives the fractal dimension as almost constant, at $D_{\text{SIS}} \approx 1.7 \pm 0.05$. In contrast, the modified SIS model results in a richer behaviour: the fractal dimension takes values in the range $[1.52 \pm 0.05; 1.85 \pm 0.05]$ and scales approximately as

$$D \left( \frac{1}{\delta} \right) = -0.154 \cdot \log_{10} \left( \frac{1}{\delta} \right).$$

This result is slightly influenced by the system size and thus allows us, in principle, to extract important information on real-world data sets.

3. WWW data analysis

As already shown in [21], the fractal dimension analysis can be performed using the time evolution of Web pages that include some keywords. Having a proper model of the processes behind the spreading process, it might be possible to obtain from the fractal dimension time evolution some quantitative measurement of important system parameters. Indeed, the fractal dimension analysis results could give insight into the underlying dynamics generating the overall signal (or the structure and social behaviour of the communities under study, if one
wants to put forward a hypothesis). For example, in our case, if the process is correctly modelled by our modified SIS model, one could calculate the memory time $\mu$ of the spreaders, from equation (7). A typical example of such analysis is presented in figure 6, where the time series, sampled every hour, is reproduced and the corresponding fractal dimension computed spans from $D = 1.2 \pm 0.2$ to $D = 1.7 \pm 0.2$. A first clear result from this analysis is that the SIS model has to be rejected, as it cannot reproduce the real data properties; indeed the real process is characterized also by fractal dimensions far away from $D_{\text{SIS}}$. On the other hand, the modified SIS model allows us to obtain a better reproduction of most of the experimental data. Indeed, working under the assumption that the main features are grasped by the modified SIS model, from the fractal analysis of the signal one can infer an average recovery time of the process. As a result, one might distinguish processes that are characterized by different timescales spanning from a few hours to a few days. Note that during the Christmas holidays the fractal dimension is increased, which signals very frequent activity and, correspondingly, from equation (7), the shortest recovery time. After that period, the fractal dimension lowers again, which means a return to usual activity.

4. Conclusions

We have shown that tools from signal processing analysis and, in particular, fractal analysis might be used to test theoretical models used to describe network dynamics. In particular, we have compared possible results from an SIS model and shown that it does not allow for a wide enough variation of the fractal dimension, thus failing to describe experimental data. We then introduced a modified SIS model to better match the real-world data, obtaining an improved description, although it is still not completely satisfactory; there is still some discrepancy between the real-world data analysis and the results from the modified SIS model introduced here. This might be due to the fact that the infection probability does not exactly scale linearly with the number of neighbours, as we have assumed. Indeed, a single-parameter model might not be sufficiently versatile for describing quantitatively complex social dynamics. For example, including contrary or multiple opinions, as shown in [31], might also improve the descriptive power of the model. Finally, the discrepancy between the real-world data analysis and the results from the modified SIS model introduced here call for further developments: an extended search for a perfect matching model and a comparison with statistically significant real-world data will be subjects of future work.
Acknowledgments

We gratefully acknowledge M Furini for support in the data harvesting and we thank the bwGRiD for computational resources.

References

[1] Barabási A-L 2011 Nature Phys. 8 14
[2] Barabási A-L and Albert R 2002 Rev. Mod. Phys. 74 47
[3] Castellano C, Fortunato S and Loreto V 2009 Rev. Mod. Phys. 81 591
[4] Leskovec J, Adamic L A and Huberman B A 2007 ACM Transactions on the Web 1 1
[5] Lewis K, Gonzalez M and Kaufman J 2012 Proc. Natl. Acad. Sci. USA 109 68
[6] Liben-Nowell D and Kleinberg J 2008 Proc. Natl. Acad. Sci. USA 105 4633
[7] Viswanath B, Mislove A, Cha M and Gummadi K P 2009 Proc. 2nd ACM Workshop on Online Social Networks pp 37–42
[8] Palazuelos C and Zorrilla M 2012 Proc. 2012 Joint EDBT/ICDT Workshops pp 9–12
[9] Lind P G, da Silva L R, Andrade J S and Herrmann H J 2007 Phys. Rev. E 76 036117
[10] Gallos L K, Barttfeld P, Havlin S, Sigman M and Mokse H A 2012 Sci. Rep. 2 454
[11] Centola D 2010 Science 329 1194
[12] Moreno Y, Nekovee M and Pacheco A F 2004 Phys. Rev. E 69 066130
[13] Peyton Young H 2011 Proc. Natl. Acad. Sci. USA 108 21285
[14] Porez-Reche F J, Ludlam J J, Taraskin S N and Gilligan C A 2011 Phys. Rev. Lett. 106 218701
[15] Shao J, Havlin S and Stanley H 2009 Phys. Rev. Lett. 103 018701
[16] Liu C and Zhang Z 2014 Comm. Nonl. Science and Numerical Simulation 19 896
[17] Araujo N A M, Andrade J S and Herrmann H J 2010 Plos One 5 e12446
[18] Moreira A A, Paula D R, Costa Filho R N and Andrade J S 2006 Phys. Rev. E 73 065101
[19] West B and Grigolini P 2010 Complex Webs: Anticipating the Improbable (Cambridge: Cambridge University Press)
[20] Montangero S and Furini M Method for analysing web data US Patent, Serial No.: 12/635.004. Furini M and Montangero S 2014 Open Journal of Web Technologies (OJWT) 1 1–14 (available at www.ronpub.com/publications/ojwt/OJWT-v1i1n01_Furini.html)
[21] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E and Makse H A 2010 Nature Phys. 6 888
[22] Bollobos B 1984 Proc. Cambridge Combinatorial Conf. in Honor of P. Erdös (Academic, 1984) p 35
[23] Carmi S, Havlin S, Kirkpatrick S, Shavitt Y and Shir E 1990 Proc. Natl. Acad. Sci. USA 104 11150
[24] Barthelemy M, Barrat A, Pastor-Satorras R and Vespignani A 2004 Phys. Rev. Lett. 92 178701
[25] Pastor-Satorras R and Vespignani A 2001 Phys. Rev. Lett. 86 3200
[26] Moreno Y and Vazquez A 2003 Eur. Phys. J. B 31 265
[27] Schrafta J A S, Ketzmerick R, Gould C, Feng Y, Kelly P J, Delage A and Wasilewski Z 1998 Phys. Rev. Lett. 80 1948
[28] Gneiting T and Schlather M 2004 SIAM Review 46 269
[29] Graphviz (www.graphviz.org)
[30] Schwarmle V, Gonzalez M C, Moreira A A, Andrade J S and Herrmann H J 2007 Phys. Rev. E 75 066108