Influence of the network topology on epidemic spreading

Ljupco Kocarev

Macedonian Academy of Sciences and Arts
University of California San Diego
University Ss Cyril and Methodius, Skopje, Macedonia
(lkocarev@ucsd.edu)

Ljubljana, May 7, 2013

Networks

- The world is networked – urban transportation systems, electric power grids, the Internet, and the Web are all large complex systems that share an important feature: *they are networked*
- Examples of networks:
  - *Social networks* – Organizational networks; Communication networks; Collaboration networks; Sexual networks
  - *Information networks* – World Wide Web: hyperlinks; Citation networks; Blog networks
  - *Technological networks* – Power grid; Airline networks; Telephone networks; Internet; Autonomous systems
  - *Biological networks* – Metabolic networks; Food webs; Neural networks; Gene regulatory networks

Networks

Network data is increasingly available:

- Large on-line computing applications where data can naturally be represented as a network:
  - On-line communities: Facebook
  - Communication: Instant Messenger
  - News and Social media: Blogging
  - Also in systems biology, health, medicine,
- Network is a set of weakly interacting entities
- Links give added value:
  - Google realized web-pages are connected
  - Collective classification

Networks

Jim Gray (1998 Turing Award address):

The emergence of “cyberspace” and the World Wide Web is like the discovery of a new continent

- Complex networks as *phenomena*, not just designed artifacts
- What are the common patterns that emerge?

Networks

Mike Steuerwalt (NSF KDI workshop, 1998):

We want Kepler's Laws of Motion for the Web

- Need statistical methods to quantify large networks
- What do we hope to achieve from models of networks?
  - Patterns and statistical propertiesof network data
  - Design principlesand models
  - Understand why networks are organized the way they are (predict behavior of networked systems)

Networks

Network science faces three general problems:

- How a network can be inferred from real data
- How to characterize the network, its structure and properties
- What the processes are that take place on networks

Social networks and viral marketing

- “The New Rules of Viral Marketing: How word-of-mouse spreads your ideas for free” by David Meerman Scott (2008)
- When 7 = 350 000 000
- When Harry Potter’s Wizarding World was first announced, only 7 people knew about this. These 7 people blogged and discussed this exciting new project to the world. Within 24 hours 350 000 000 have already heard of Harry Potter’s Wizarding World at Universal Studios.

Epidemic models

- The earliest account of mathematical modeling of spread of disease was carried out in 1766 by Daniel Bernoulli.
- A. G. McKendrick and W. O. Kermack: A Contribution to the Mathematical Theory of Epidemics (1927)
- Reed-Frost epidemic model (1928) – one of the simplest stochastic epidemic models

SIS model

- Population of N individuals, connected in a network structure represented by a graph $G = (V, E)$ with node set V and edge set E
- Each node can be in one of two possible states: susceptible (S) and infective (I)
- $\mathbf{s}_i(t) = [s_i^S(t) \;\; s_i^I(t)]^T$ – status vector, an indicator vector containing a single 1 in the position corresponding to the present state, and 0 everywhere else
- $\mathbf{p}_i(t) = [p_i^S(t) \;\; p_i^I(t)]^T$ – probability mass-function (PMF) of node i at time t: $p_i^S(t) + p_i^I(t) = 1$

SIS model

The evolution of SIS is described by the following equations:

$$\begin{aligned} p_i^I(t+1) &= s_i^S(t) f_i(t) + (1-\delta) s_i^I(t), \\ \mathbf{s}_i(t+1) &= \mathit{MultiRealize}[\mathbf{p}_i(t+1)]. \end{aligned}$$

- *MultiRealize*$[\cdot]$ – performs a random realization for the PMF given with $\mathbf{p}_i(t+1)$
- The first term on the right hand side is the probability that a susceptible node i is infected $f_i(t)$ by at least a neighbor
- The second term stands for the probability that infected node i at time t does not recover
- $0 \leq \delta \leq 1$ – the cure rate of the virus

SIS model

$$f_i(t) \quad = \quad 1 - \prod_{j=1}^{N} \left[1 - \beta r_{ij} s_j^I(t) \right] .$$

β – probability of disease transmission from an I node to an S node

- Contact process and reactive process
- The contact process is a dynamical process that involves a single stochastic contagion per infective node per unit time.
- The reactive process there are as many stochastic contagions per unit time as there are neighbours to a node.

SIS model

- The distinction between the two processes is reflected in the probability $f_i(t)$ that a susceptible node i receives the infection from any combination of its infective neighbours.
- $r_{ij} = a_{ij} / \sum_j a_{ij}$ – contact process
- $r_{ij} = a_{ij}$ – reactive process

SIR model

The evolution of SIR is described by the following equations:

$$
\begin{aligned}
p_i^I(t+1) &= s_i^S(t)\left[1-\prod_{j=1}^{N}\left[1-\beta r_{ij}s_j^I(t)\right]\right], \\
p_i^R(t+1) &= s_i^I(t)+s_i^I(t) \\
\mathbf{s}_i(t+1) &= \mathit{MultiRealize}[\mathbf{p}_i(t+1)].
\end{aligned}
$$

$$
p_i^S(t+1)+p_i^I(t+1)+p_i^R(t+1)=1
$$

Spreading processes on networks

Several approaches to study processes on networks:

- Mathematics (stochastic, deterministic, dynamical systems approach)
- Physics (statistical physics, the theory of phase transitions and critical phenomena)
- Computer science (optimal solutions, computational complexity theory)

The problem of modeling how diseases spread among individuals has been intensively studied for many years. Today the problem has attracted a lot of interest in a view of possible applications in social networks and viral marketing.

Epidemic models in social networks

D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network. Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2003.

- When node *i* first becomes active in step *t*, it is given a single chance to activate each currently inactive neighbor *j*; it succeeds with a probability $\beta_{i,j}$ $(= \beta)$
- If *i* succeeds, then *j* will become active in step $t + 1$; but whether or not *i* succeeds, it cannot make any further attempts to activate *j* in subsequent rounds.
- The process runs until no more activations are possible.

Epidemic models in social networks

- Influence of a set of nodes A, denoted $\sigma(A)$, – expected number of active nodes at the end of the process, given that A is an initial active set
- The influence maximization problem – for a parameter k, find a k-node set of maximum influence
- NP-hard problem
- Natural greedy strategy obtains a solution that is provably within 63% of optimal (for several classes of models)
- A general approach for reasoning about the performance guarantees of algorithms for influence problems in social networks

Thresholds in epidemic models

M. Draief, A.Ganesh, L. Massoulie: “Thresholds for virus spread on networks”, Annals of Applied Probability, Vol. 18, No. 2 (2008), pp 359–378

- Suppose $\beta\lambda_{1,A} < 1$. Then, the expected number of removed nodes satisfies

$$N_R(\infty) \leq \frac{1}{1-\beta\lambda_{1,A}}\sqrt{nN_I(0)}$$

- $\lambda_{1,A}$ – the largest eigenvalue of the adjacency matrix
- $N_I(0)$ – number of initial infectives

Deterministic model

Deterministic model ($x_i = p_i^I$):

$$\begin{aligned} x_i(t+1) &= [1 - x_i(t)]\, f_i(t) + (1-\delta) x_i(t) \\ f_i(t) &= 1 - \prod_{j=1}^{N} \left[1 - \beta a_{ij} x_j(t)\right]. \end{aligned}$$

- The origin $x_i = 0$ $(\forall i)$ is a fixed point of the system
- The origin is stable when $1 - \delta + \beta\lambda_{1,A} < 1$, where $\lambda_{1,A}$ is the largest eigenvalue of the adjacency matrix
- $\beta/\delta > 1/\lambda_{1,A}$ the disease will reach an endemic state.

Central questions

What is the influence of the graph topology on virus (disease) spreading, more precisely, on the probability that given note is infective?

How local structural properties of the network constrain the interval of possible values of the probability that given note is infective?

Local versus global properties

V. M. Preciado, M. Draief, A. Jadbabaie, “Structural Analysis of Viral Spreading Processes in Social and Communication Networks Using Egonets”, Sep 2012, arXiv:1209.0341

A mathematical framework, based on algebraic graph theory and convex optimization, is proposed, to study how local structural properties of the network constrain the interval of possible values in which the largest eigenvalue must lie.

Local versus global properties

V. M. Preciado, A. Jadbabaie, G. C. Verghese,“Structural Analysis of Laplacian Spectral Properties of Large-Scale Networks”, Sept 2012, IEEE Automatic Control, accepted for publication, arXiv:1107.5676

Laplacian spectral moments and spectral radius are strongly constrained by local structural features of the network; however, local structural features are not enough to estimate the Laplacian spectral gap.

Local versus global properties

R. Agarwal, M. Caesar, P. B. Godfrey, B. Y. Zhao, “ Shortest Paths in Less Than a Millisecond”, SIGCOMM WOSN 2012, arXiv:1206.1134

By using the idea of vicinity intersection, where the vicinity of a user is a (carefully defined) subset of users in its neighborhood, the problem of answering point-to-point shortest path queries on massive social networks is reduced to milliseconds. For instance, for the LiveJournal social network (roughly 5 million nodes and 69 million edges), the technique can answer 99.9% of the queries in less than a millisecond.

p-hop subgraph

- Let i be arbitrary node of the graph $G = (V, E)$, $i \in V$, and let $n_i = \max_x l(i, x)$. Let $V_i^0 = \{i\}$.
- We define a subgraph $G_i^p = (V_i^p, E_i^p)$ of $G = (V, E)$ as follows:

$$\begin{aligned} V_i^p &= \{x | x \in V, 0 \leq l(i, x) \leq p\} \\ E_i^p &= \{xy | xy \in E, x \in V_i^p, y \in V_i^{p-1}\}, \end{aligned}$$

  where $p = 1, \ldots, n_i + 1$.
- We say that G_i^p is a p-hop subgraph of G extracted by starting at node i.

p-hop subgraph

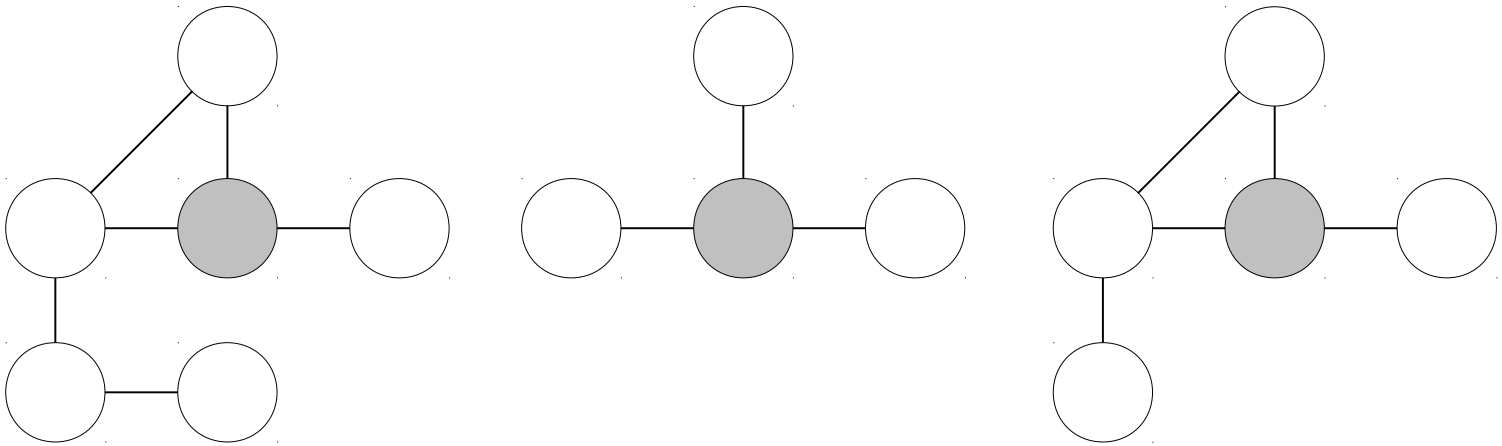


Figure: b) and c) 1-hop and 2-hop subgraph of the graph shown in a) extracted by starting at the gray node.

Upper and lower bounds

- We will prove that

$$l_i^1 < l_i^2 < \ldots < l_i^p \ldots \leq x_i^* < \ldots u_i^p < \ldots < u_1^2 < u_i^1$$

- The bounds l_i^1 and u_i^1 are obtained by considering only (first) neighbors of i.
- The bound u_i^1 depends on the degree of the node i, that is, the information contained in the 1-hop subgraph of G extracted by starting at node i, while for the bound l_i^1 one computes the SIS model on the subgraph G_i^1, which is the subgraph of neighbors of i.

Upper and lower bounds

- The bounds l_i^2 and u_i^2 are obtained by considering second neighbors of i (neighbors of the first neighbors).
- l_i^2 and u_i^2 reflect the topology of 2-hop subgraph of G extracted by starting at node i.
- $$d_i^p = u_i^p - l_i^p \qquad \Delta\rho_p = \sum_{i=1}^{N} \frac{d_i^p}{N},$$

Example: real-world e-mail network

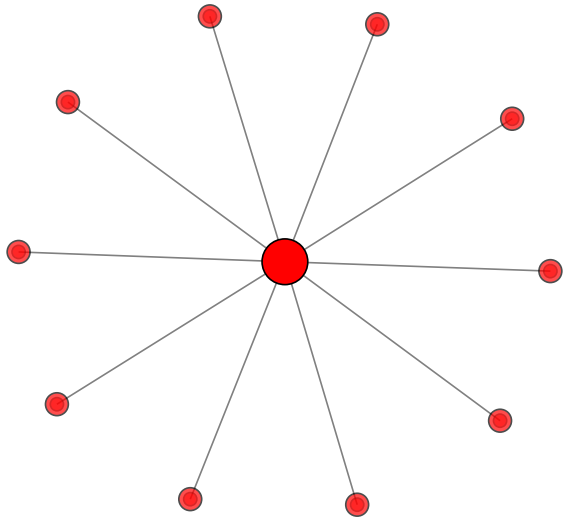


Figure: 1-hop neighborhood for a node extracted from a real-world e-mail network with 33696 nodes. The node (largest in size) has 10 direct neighbors (medium sized). The probability of infection for the given node obtained after simulating a particular configuration of the SIS model was 0.373. The probability of infection for the node given the 1-hop neighborhood (node's degree) is calculated to be between 0.006 and 0.503.

Example: real-world e-mail network

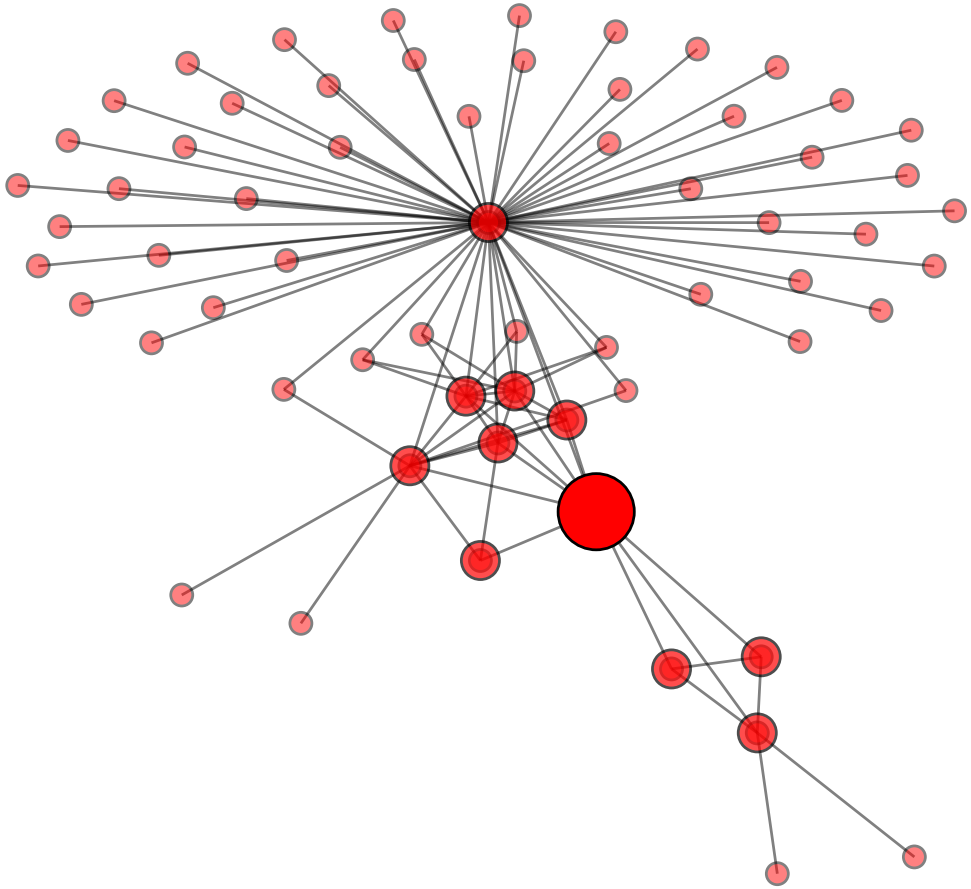


Figure: The 2-hop neighborhood contains 62 nodes and 92 edges. Peripheral nodes are smallest in size and are two hops away from the central node. The probability of infection for the node given the 2-hop neighborhood topology is calculated to be between 0.297 and 0.416.

Reactive process

Theorem

Let Φ be the family of all possible simple and connected graphs $G = (V, E)$ with $|V| \geq 2$. Let $\mathbf{x}(G)^* = [x_1^* x_2^* \ldots x_N^*]$ be the stationary solution different from the origin and let i be arbitrary node of the graph $G = (V, E)$, $i \in V$. Let

$$u_i^n = \frac{1 - \prod_{j=1}^{N}(1 - \beta a_{ij} u_j^{n-1})}{1 - \prod_{j=1}^{N}(1 - \beta a_{ij} u_j^{n-1}) + \delta}$$

where $u_i^0 = 1/(1+\delta)$. Then for all i, x_i^* is bounded by

$$x_i^* < \ldots < u_i^n < \ldots < u_i^1 < u_i^0$$

Reactive process

Theorem

Consider arbitrary node i of the graph $G = (V, E)$ and let $G_i^p = (V_i^p, E_i^p)$ be the p-hop subgraph of G extracted by starting at node i. Write $n = \left|V_i^p\right|$. Let $\mathbf{x}(G)^* = [x_1^* x_2^* \dots x_N^*]$ and $\mathbf{x}(G_i^p)^* = [I_1^p I_2^p \dots I_n^p]$ be the stationary solution different from the origin for the graphs $G = (V, E)$ and $G_i^p = (V_i^p, E_i^p)$, respectively. Then x_i^* is bounded by

$$I_i^1 < I_i^2 < \dots < I_i^{n_i} < I_i^{n_i+1} = x_i^*$$

for all $i \in V$.

Theorem

Let $\mathbf{x}(G)^* = [x_1^* x_2^* \dots x_N^*]$ *be the stationary solution different from the origin and let*

$$I_i^n = \frac{1 - \prod_{j=1}^{N}(1 - \beta r_{ij} I_j^{n-1})}{1 - \prod_{j=1}^{N}(1 - \beta r_{ij} I_j^{n-1}) + \delta}$$

and

$$u_i^n = \frac{1 - \prod_{j=1}^{N}(1 - \beta r_{ij} u_j^{n-1})}{1 - \prod_{j=1}^{N}(1 - \beta r_{ij} u_j^{n-1}) + \delta}$$

Theorem

In the last equations:

$$I_i^0 = \frac{1 - e^{-\beta x_i^*}}{1 - e^{-\beta x_i^*} + \delta} \text{ and } u_i^0 = 1 - \frac{\delta}{\beta}$$

then x_i^* *is bounded by*

$$I_i^0 < I_i^1 < \ldots < I_i^n \ldots < x_i^* \ldots \leq u_i^n \leq \ldots \leq u_i^1 \leq u_i^0$$

for all i.

Average size of p-hop neighborhood

Table: Average size of p-hop neighborhood for the Enron e-mail network. $|E^p|$ is an average of $|E_i^p|$ over all nodes i and $|E|$ is the total number of edges in the network.

p	$\|E^p\|$	$\|E^p\| / \|E\|$
1	10	0.0003
2	1538	0.004
3	45067	0.125
4	207496	0.574

Reactive process

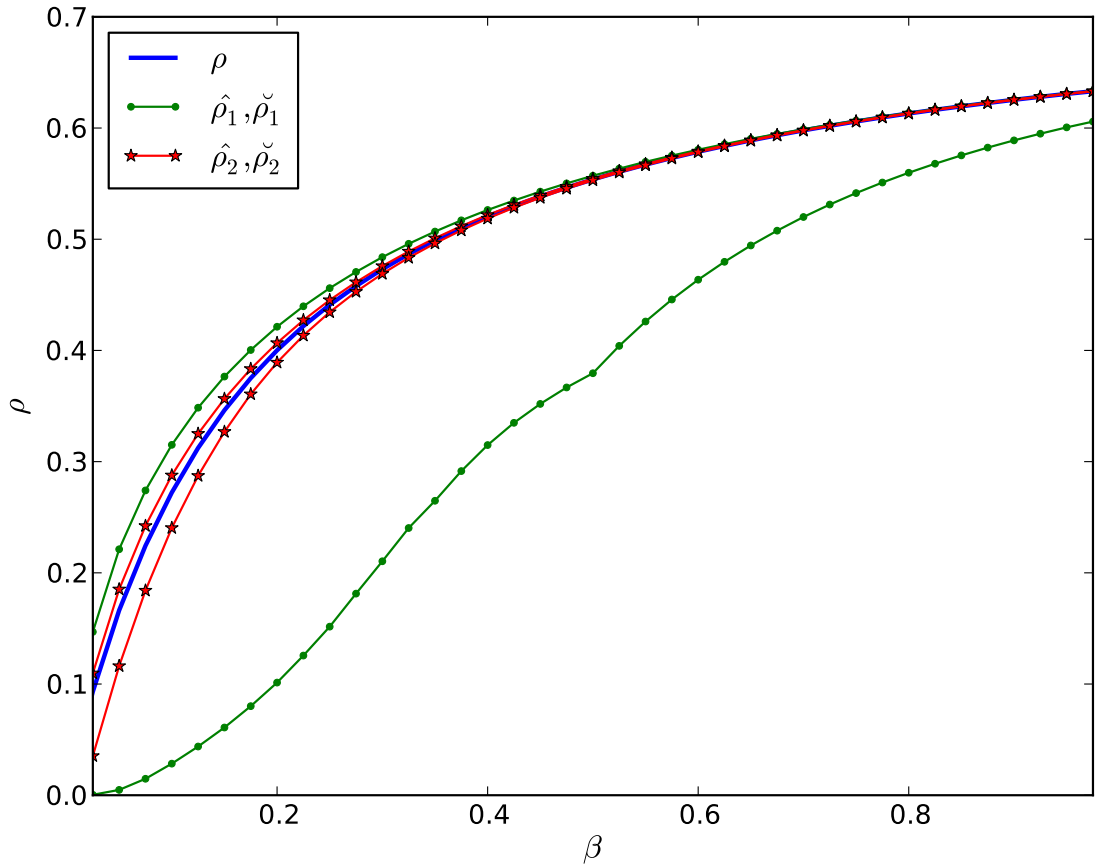
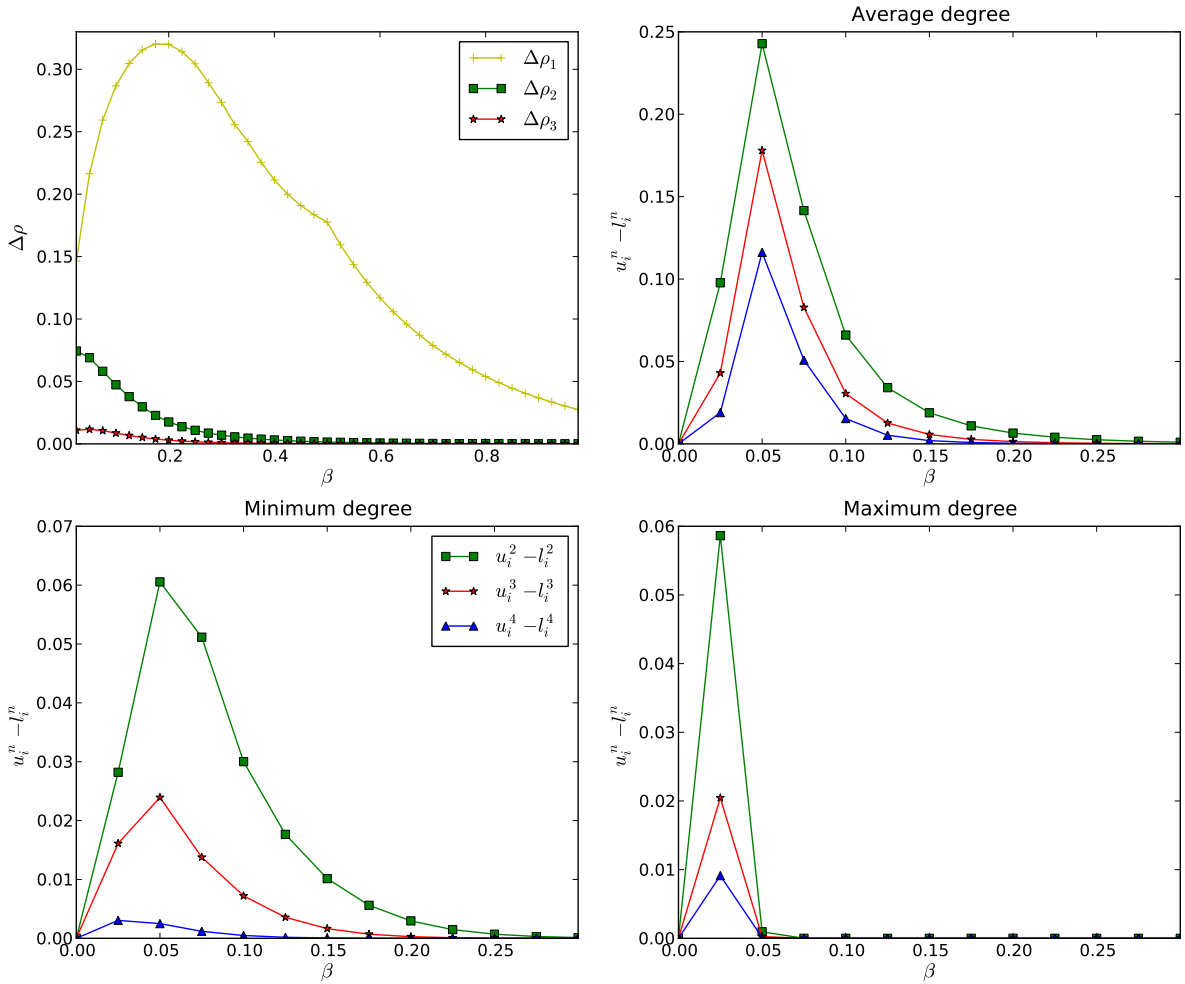


Figure: The density of infective nodes in the Enron e-mail network as the transmission parameter β is varied, and $\delta = 0.5$, along with the upper and lower bounds, $\hat{\rho}_p$ and $\breve{\rho}_p$, on ρ using 1-hop and 2-hop topology information.

Reactive process



Contact process

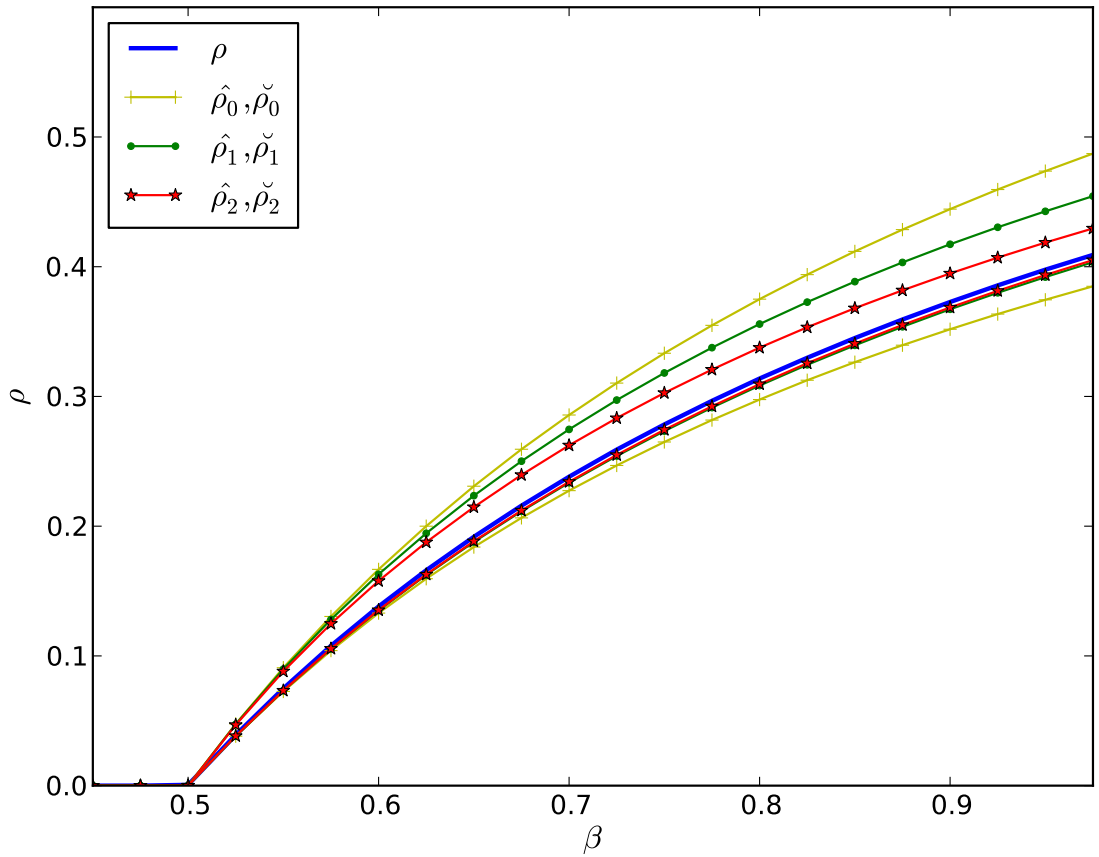
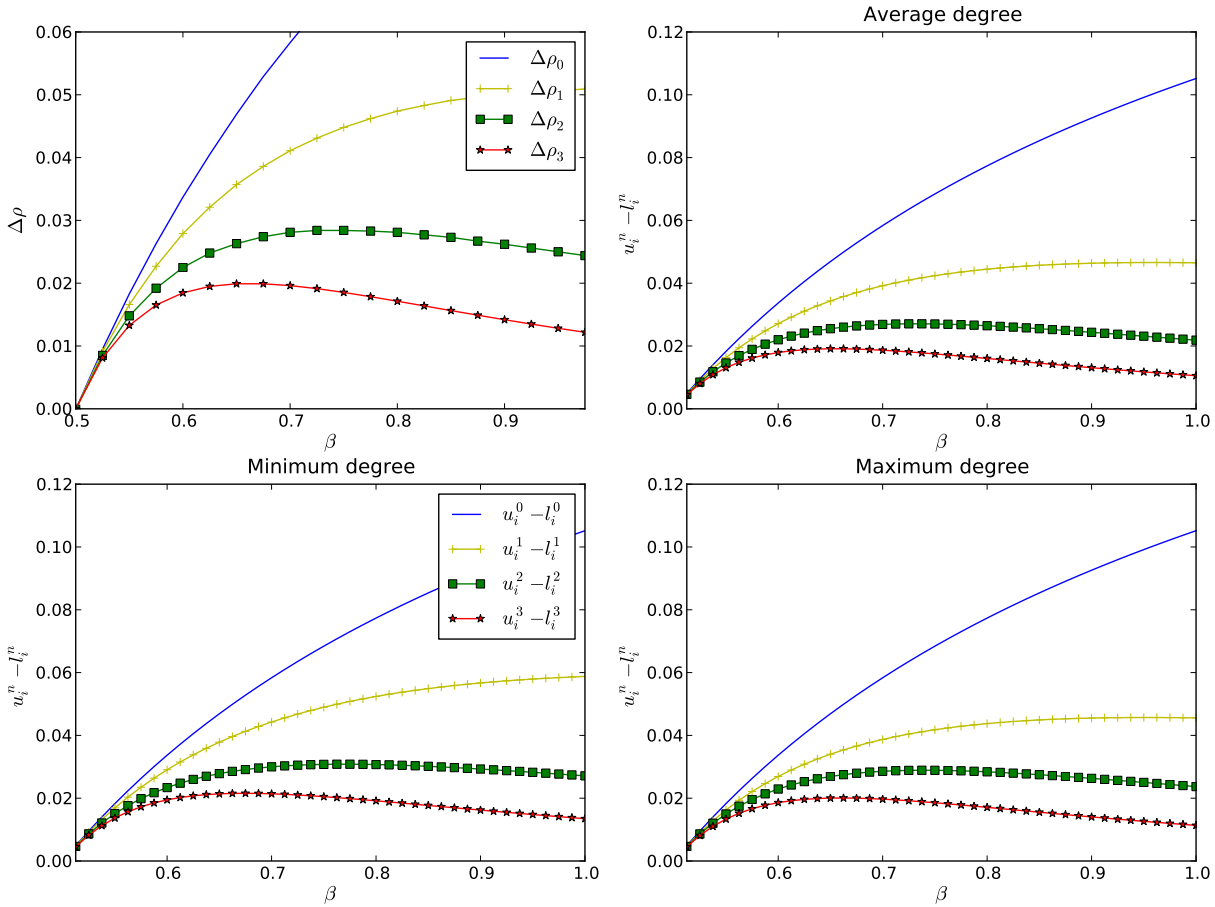


Figure: The expected infection density ρ in the endemic state for the Enron e-mail network for the contact process as β is varied, and $\delta = 0.5$. The bounds on ρ calculated with no topology information $\hat{\rho}_0$, $\breve{\rho}_0$, 1-hop topology information $\hat{\rho}_1$, $\breve{\rho}_1$, and 2-hop topology information $\hat{\rho}_2$, $\breve{\rho}_2$ are depicted as well.

Contact process



Summary

- One can estimate the probability of being infective using only local information (considering only $n-$hop local topology, for small n), without knowing the whole network.
- From this local information one can also estimate the density of being infective on the whole network, as well as assess the extend to which the topology affects the outcome of the infection on macroscopic level.
- The results are extendable to other ergodic models (such as SIRS, for example) and are related to all types of spreading (idea, failure, rumor).

References

- D. Trpevski, W. K. S. Tang, and L. Kocarev, Model for rumor spreading over networks, Physical Review E 81, 056102 (11 pages) 2010
- D. Smilkov and L. Kocarev, Influence of the network topology on epidemic spreading, Physical Review E 85, 016114 (10 pages), 2012
- I. Tomovski and L. Kocarev, Simple Algorithm for Virus Spreading Control on Complex Networks, IEEE Transactions on Circuits and Systems I: Regular Papers, Volume: 59, Issue: 4, Page(s): 763 – 771, 2012

Thanks

Thanks to:

- D. Smilkov (Boston), I. Tomovski and D. Trpevski (Skopje), W. K. S. Tang (Honk Kong)
- ONR: “Optimization and Performance Enhancement of Complex Networks using Sensors” (N62909-10-1-7074)
- MON: “Annotated graphs in System Biology”