Analytically tractable processes on networks

Ljupco Kocarev

University of California San Diego

CERTH, 25 May 2011

Outline

1. Motivation
   - Networks
   - Random walk and Consensus
   - Epidemic models
   - Spreading processes on networks
2. Our Contribution
   - Linear processes on networks
   - Number of infective nodes in epidemic models
   - Topology independent spreading processes

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*

- 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

Processes on Networks

As many dynamic phenomena as networks:

- biologists study reaction kinetics on metabolic networks
- computer scientists monitor the flow of information on computer networks
- epidemiologists, sociologists, and economists explore the spread of viruses and ideas on social networks
- electrical engineers study and control power grids

Two basic problems that we have studied:

- Linear processes on networks
- Epidemic models

Random walk and consensus

The simplest dynamical processes on networks are linear processes:

$$x_i(t+1) = \sum_j b_{ij} x_j(t),$$

- x_i – quantity associated to node i
- $B = [b_{ij}]$ – a matrix related to the adjacency matrix A
- $\mathbf{x} = [x_1, \ldots, x_N]^T$ – column vector of length N

$$\mathbf{x}(t+1) = B\mathbf{x}(t),$$

Random walk and consensus

- Random walk: B is column stochastic, $b_{ij} = a_{ij}/s_j$
- Consensus: B is row stochastic, $b_{ij} = a_{ij}/s_i$
- $s_i = \sum_j a_{ij}$ – the degree of node i

$$\mathbf{x}(t) = B^t\mathbf{x}(0) \rightarrow \begin{cases} (\pi \otimes \mathbf{1}_N^T)\mathbf{x}(0) & \text{random walk} \\ (\pi^T \otimes \mathbf{1}_N)\mathbf{x}(0) & \text{consensus} \end{cases}$$

$$= \begin{cases} \pi \mathbf{1}_N^T \mathbf{x}(0) = \pi \, \|\mathbf{x}(0)\| \\ \mathbf{1}_N \pi^T \mathbf{x}(0) \end{cases}$$

- π – dominant eigenvector of B
- $\mathbf{1}_N$ – length N column vector of 1
- $C \otimes D$ – Kronecker product

Random walk and consensus

- Random walks – dynamical processes aiming at modeling the diffusion of some quantity or information on networks
- Random walks have a long history is physics, chemistry, biology, computer science (PageRank, BLAST), and so on

- In networks of agents, consensus means to reach an agreement regarding a certain quantity of interest that depends on the state of all agents
- Consensus problems have a long history in physics (synchronization), electrical and control engineering, computer science (distributed computing), and so on

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:
  - Each infective individual at time t independently makes contacts with all other individuals in the population with some probability p, and if a contacted individual is susceptible, it becomes infected at time $t + 1$
  - At time $t + 1$, the infective individuals from time t are removed from the epidemic process

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:

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

- $\mathit{MultiRealize}[\cdot]$ – 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

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

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.

Deterministic models

- Deterministic models
  - Linear models

$$\mathbf{x}_i(t+1) = \sum_{j=1}^{N} b_{ij} D_{ij} \mathbf{x}_j(t)$$

  - Epidemic models

$$x_i(t+1) = [1 - x_i(t)] \left[1 - \prod_{j=1}^{N} [1 - \beta a_{ij} x_j(t)] \right] + (1 - \delta) x_i(t)$$

$$x_i(t+1) = [1 - x_i(t)] \left[\sum_{j=1}^{N} \beta b_{ij} x_j(t) \right] + (1 - \delta) x_i(t)$$

- Non-trivial solutions
- Stability analysis

Linear processes on networks – summary

- Broad class of analytically solvable processes on networks
- Random walk and consensus process
- The model is analytically solvable:
  - dynamical equation for each node may be different
  - the network may have an arbitrary finite graph and influence structure
- In the homogeneous case the model is decomposable:
  - equilibrium behavior can be expressed as an explicit function of network topology and node dynamics

Epidemic models – summary

- The simplest ergodic epidemic model:
  susceptible – infective – susceptible (SIS)
- All results are derived for the SIS model but can be extended to all other models
- The presented results are related to all types of spreading (idea, failure, rumor), regardless on the type of the spread agent

Outline

- Linear processes on networks
  - Homogeneous processes
  - Heterogeneous processes
  - Network hierarchy
- Epidemic models
  - Number of infective nodes in epidemic models
  - Topology independent spreading processes

Linear processes on networks

- $\mathbf{x}_i = [x_i^1 \; x_i^2 \ldots x_i^{m_i}]^T$ – a nonnegative m_i-dimensional column vector
- B – a stochastic $N \times N$ matrix related to the adjacency matrix A
- D_{ij} – an $m_i \times m_j$ nonnegative matrix such that each row (column) of D_{ij} sums up to 1

The evolution of each node variables has the following form:

$$\mathbf{x}_i(t+1) \quad = \quad \sum_{j=1}^{N} b_{ij} D_{ij} \mathbf{x}_j(t),$$

for all $i = 1, \ldots N$.

Two approaches

1. Consider each node i as a complex system: node i is described with a a vector of quantities (not a scalar quantity)
2. Consider a network with N nodes: each node i actually being a network with m_i internal nodes (the total number of nodes in this network of networks is $m_1 + \ldots + m_N$)

Network of Markov chains

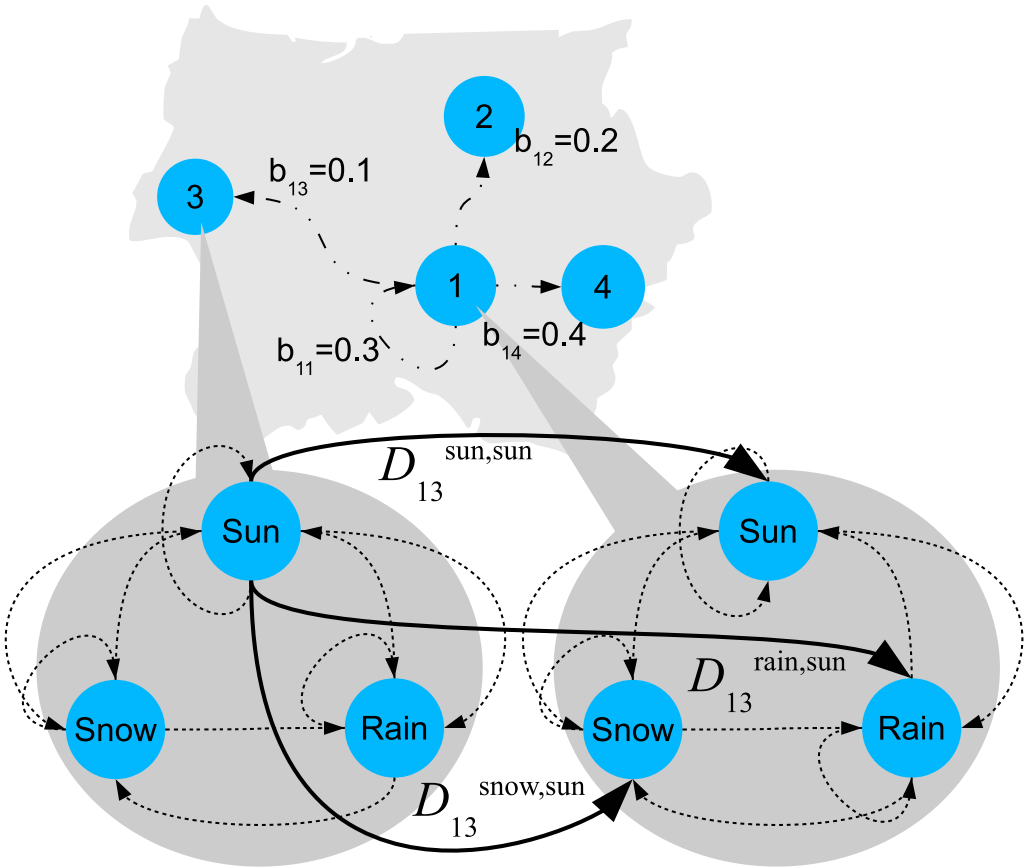

Figure: Network of Markov chains describing weather dynamics.

Random walk in a network of networks

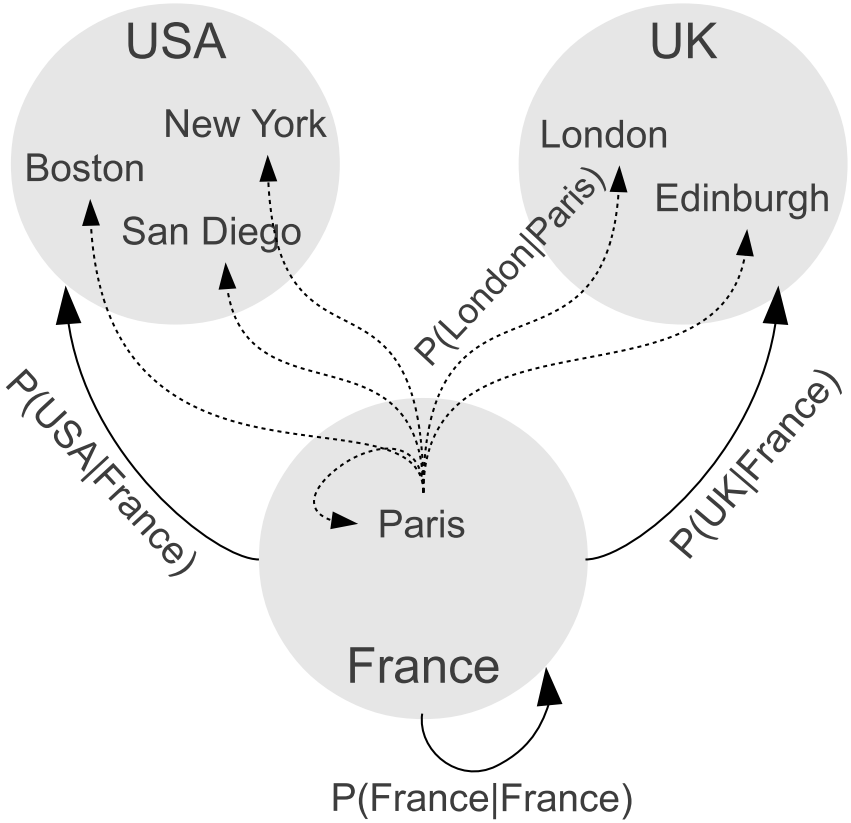

Figure: Random walk in a network of networks: a walker makes a 2-step decision for where to go; it first chooses a country, then a city within that country.

Homogeneous processes

- $m_i = m$ and $D_{ij} = D \neq I_m$ (I_m is $m \times m$ identity matrix)
- homogeneous processes – the local dynamics in each node is described with the same evolution equation

The evolution of each node variables has the following form:

$$\mathbf{x}_i(t+1) \quad = \quad (B \otimes D)\,\mathbf{x}(t) \equiv H\mathbf{x}(t)$$

Random walk

- B and D are column stochastic
- The stationary solution of the random walk is

$$\begin{aligned} \mathbf{x}(t) &= H^t \mathbf{x}(0) \to (\pi \otimes \rho) \, \|\mathbf{x}(0)\| \\ \lim_{t \to \infty} x_i^k(t) &= \|\mathbf{x}(0)\| \, \pi_i \rho_k \end{aligned}$$

for all $i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, m$.

Consensus

- B and D are row stochastic
- The model stationary solution is

$$\begin{aligned} \mathbf{x}(t) &= H^t \mathbf{x}(0) \to \mathbf{1}_{Nm}(\pi^T \otimes \rho^T)\mathbf{x}(0) \\ \lim_{t\to\infty} x_i^k(t) &= (\pi^T \otimes \rho^T)\mathbf{x}(0) \end{aligned}$$

for all $i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, m$.

Network of identical Markov chains

- B is row stochastic and D is column stochastic
- The model stationary solution is

$$\begin{aligned} \mathbf{x}(t) &= H^t \mathbf{x}(0) \to (\pi_1 \|\mathbf{x}_1(0)\| \dots \pi_N \|\mathbf{x}_N(0)\|)[\rho \dots \rho]^T \\ \lim_{t\to\infty} x_i^k(t) &= (\pi_1 \|\mathbf{x}_1(0)\| \dots \pi_N \|\mathbf{x}_N(0)\|)\rho_k. \end{aligned}$$

- If $\|\mathbf{x}_i(0)\| = c, \forall i = 1, \dots, N$, then the model satisfies the consistency rule $\|\mathbf{x}_i(t+1)\| = \|\mathbf{x}_i(t)\| = c$.
- For $c = 1$, the process corresponds to N Markov chains:

$$\lim_{t\to\infty} x_i^k(t) = \rho_k.$$

- *In a network of N identical Markov chains, the equilibrium solution does not depend on the graph topology*

Heterogeneous processes

Local dynamics differs for each node in the network

$H = B \otimes \{D_{ij}\} \equiv [H_{ij}], \quad H_{ij} = b_{ij} D_{ij}$

$$\begin{aligned} \mathbf{x}(t+1) &= B \otimes \{D_{ij}\}\mathbf{x}(t) = H\mathbf{x}(t), \\ \mathbf{y}(t+1) &= H\mathbf{y}(t) \end{aligned}$$

$$\mathbf{y} = [\underbrace{y_1 y_2 \dots y_{m_1}}_{\mathbf{x}_1} \underbrace{y_{m_1+1} y_{m_1+2} \dots y_{m_1+m_2}}_{\mathbf{x}_2} \dots y_s]^T$$

$\mathbf{y}$ – column vector of length $s = m_1 + \ldots + m_N$

H – $s \times s$ matrix

$\gamma = [\gamma_1, \gamma_2, \ldots \gamma_s]^T$ – dominant eigenvector of H

Random walk

B is column stochastic; each column of D_{ij} sums up to 1

- The model satisfies the consistency rule
  $\|\mathbf{x}(t+1)\| = \|\mathbf{x}(t)\|$
- H is column stochastic matrix
- Assuming that H is irreducible matrix, one can show

$$\lim_{t\to\infty} y_i(t) \quad = \quad \|\mathbf{x}(0)\|\,\gamma_i$$

for all $i = 1, 2, \ldots, s$

Consensus

B is row stochastic; each row of D_{ij} sums up to 1

- H is row stochastic
- Assuming that H is irreducible matrix, one can show

$$\lim_{t\to\infty} y_i(t) = \sum_i \gamma_i y_i(0)$$

for all $i = 1, 2, \ldots, s$.

Network of different Markov chains

B is row stochastic; each column of D_{ij} sums up to 1

- The matrix H is not stochastic, however 1 is its dominant eigenvalue
- The multiplicity of 1 is tied to the structure of the underlying graph
- Assuming that H is irreducible matrix, one can show
  - The consistency rule is satisfied at local level: at each node $\|\mathbf{x_i}(t+1)\| = \|\mathbf{x_i}(t)\|$.
  - Let α_i be a stationary distribution of the Markov chain at node i:

$$\alpha_i = [\gamma_{a+1} \dots \gamma_{a+m_i}]^T, a = \sum_{j=1}^{i-1} m_j.$$

Network hierarchy

- Complex networks exhibit hierarchical organization – the network self-organizes into modules that further subdivide into modules of modules, and so forth over multiple scales
- The first problem – developing a general framework to study network hierarchy taking into account the processes on networks

Hierarchical processes:

$$\mathbf{x}(t+1) = \left(\ldots \left(B \otimes \{D_{ij}^1\} \right) \ldots \otimes \{D_{ij}^h\} \right) \mathbf{x}(t),$$

$H_1 = B \otimes \{D_{ij}^1\}$, $H_2 = H_1 \otimes \{D_{ij}^2\}$, so on and $H_h = H_{h-1} \otimes \{D_{ij}^h\}$

Network hierarchy

- The second problem: decomposing a given graph into subgraphs (the matrix H into matrices B and D_{ij})
- The decomposition of H has several advantages

1. Random walk
   - Stationary solution is one of the most used centrality measures in networks
   - The decomposition can be used to obtain a high-level view of stationary dynamics by lumping nodes into super-nodes
   - This reduces the size of the system, and thus, the time it takes to compute the solution
2. Consensus
   - High-level stationary solution can be used to obtain approximation for the stationary consensus values

Stochastic SIS model

$$\begin{aligned} p_i^I(t+1) &= s_i^S(t) f_i(t) + (1-\delta) s_i^I(t) \\ f_i(t) &= 1 - \prod_{j=1}^{N} \left[1 - \beta a_{ij} s_j^I(t)\right]. \end{aligned}$$

- β – probability of disease transmission from an I node to an S node
- Let $s_i^S(t) = 1$ and let $N(i;t)$ be a set of all infected neighbors of i at time t. Then

$$p_i^I(t+1) = 1 - \prod_{j \in N(i;t)} [1-\beta]$$

is the probability that the node i changes its status from S to I at time $t+1$.

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. Let $a_i(G) = x_i(\infty)$ be the unique globally stable fixed point different from the origin for the graph G

Number of infective nodes

Following conclusion holds in arbitrary graphs:

- The probability of node i to be in state I when $t \to \infty$, $a_i(G)$ is bounded:

$$a_i(G) \leq \frac{1}{1+\delta}$$

Moreover,

$$N_I = \sum_i p_i^I \equiv \sum_i a_i(G) \leq \sum_i \frac{1}{1+\delta} = \frac{N}{1+\delta}$$

Number of infective nodes

- The distribution of the number of infective nodes (in the equilibrium state – when t goes to infinity) has two peaks:

$$a_{hub} \equiv \frac{1}{1+\delta} \qquad a_{leaf} \equiv \frac{\beta}{\beta + \delta(1+\delta)}$$

- The values a_{hub} and a_{leaf} do not depend on network topology.
- The fraction of nodes that behave as hubs and leaves is always comparable with the total number of the nodes in the network (for some values of β).

Arbitrary graphs

Let G_N be a family of graphs for which the maximum node degree k_{max} is unbounded when $N \to \infty$. Then as $N \to \infty$:

- The critical value of β, β_{cr}, for which the point $a_i(G_N)$ is a stable fixed point, tends to zero .
- For given $\beta \geq \beta_{cr}$ and δ, and arbitrary $\epsilon > 0$, one can find a degree value k_c such that $a_{hub} - \epsilon < a_i(G_N) < a_{hub}$ holds for all nodes i for which $k_i \geq k_c$.
- The threshold k_c depends on δ (when ϵ and β are fixed). As $\delta \to 0$, the SIS model approaches the SI model, and the approximation $a_i(G_N) \approx a_{hub}$ becomes more accurate and holds for smaller values of k_c. When $\delta \to 1$ the approximation $a_i(G_N) \approx a_{hub}$ becomes more inaccurate and holds for larger values of k_c.

Real networks

- We study 12 real networks, all of them as undirected graphs considering only their giant components:

  http://snap.stanford.edu/

- As a typical example we present results only for the Enron e-mail network in more detail.
- The giant component of the Enron network has $N = 33696$ nodes; $k_{max} = 1383$ and $\lambda_1 = 125.906$.
- Epidemic occurs for $\beta \geq \beta_{cr} = 0.003971$ when $\delta = 0.5$.

Enron network

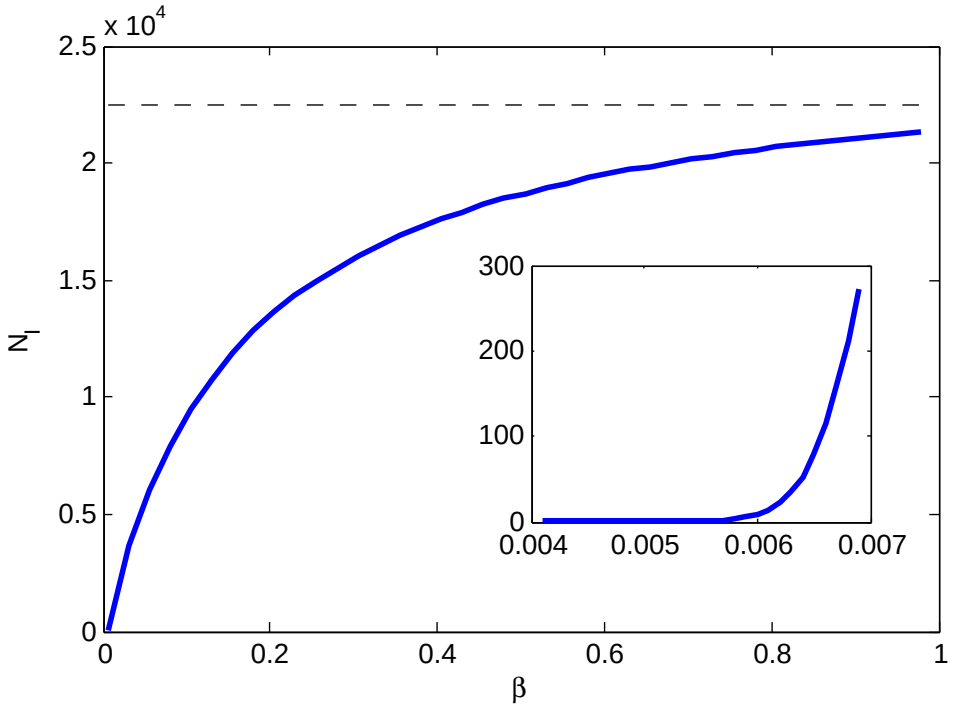

Figure: The number of infective nodes N_I in the endemic state for the Enron network as β is varied, $\delta = 0.5$. The inset shows N_I for values of β slightly above the epidemic threshold. Dashed line is the upper bound on N_I.

Enron network

Since real networks are finite, the inequality $a_{hub} - \epsilon < a_i < a_{hub}$ does not hold for arbitrarily small values of ϵ.

- For given $\beta \geq \beta_{cr}$ and δ, and arbitrary $\epsilon > 0$, one can find a degree value k_c such that $a_{hub} - \epsilon < a_i(G_N) < a_{hub}$ holds for all nodes i for which $k_i \geq k_c$.

- $\delta = 0.5$ and $\epsilon = 0.01$ – we calculate k_c for different values of β: $k_c = 187$ for $\beta = 0.1$, $k_c = 51$ for $\beta = 0.2$, and so on until $k_c = 6$ for $\beta = 0.8$.

Enron network

- When ϵ and β are fixed, the threshold k_c depends only on δ. As $\delta \to 0$, the SIS model approaches the SI model, and the approximation $a_i(G_N) \approx a_{hub}$ becomes more accurate and holds for smaller values of k_c. When $\delta \to 1$ the approximation $a_i(G_N) \approx a_{hub}$ becomes more inaccurate and holds for larger values of k_c.

- $\beta = 0.1$ and $\epsilon = 0.01$: for $\delta = 0.01$ we obtain $k_c = 8$, for $\delta = 0.5$, $k_c = 187$, and for $\delta = 0.99$, $k_c = 1384$.

Enron network

- Let $\epsilon_{cr} = \epsilon_{cr}(\delta, \beta)$ be the critical value of ϵ for which, given δ and β, the inequality $a_{hub} - \epsilon < a_i < a_{hub}$ holds only for nodes with maximum degree $k_i = k_{max}$.

- For the values $\beta = 0.004 > 0.0039$, $\beta = 0.0376$, and $\beta = 0.1$, we obtain $\epsilon_{cr} = 0.6666667$, $\epsilon_{cr} = 0.01$, and $\epsilon_{cr} = 2.69 \times 10^{-10}$, respectively.

Enron network

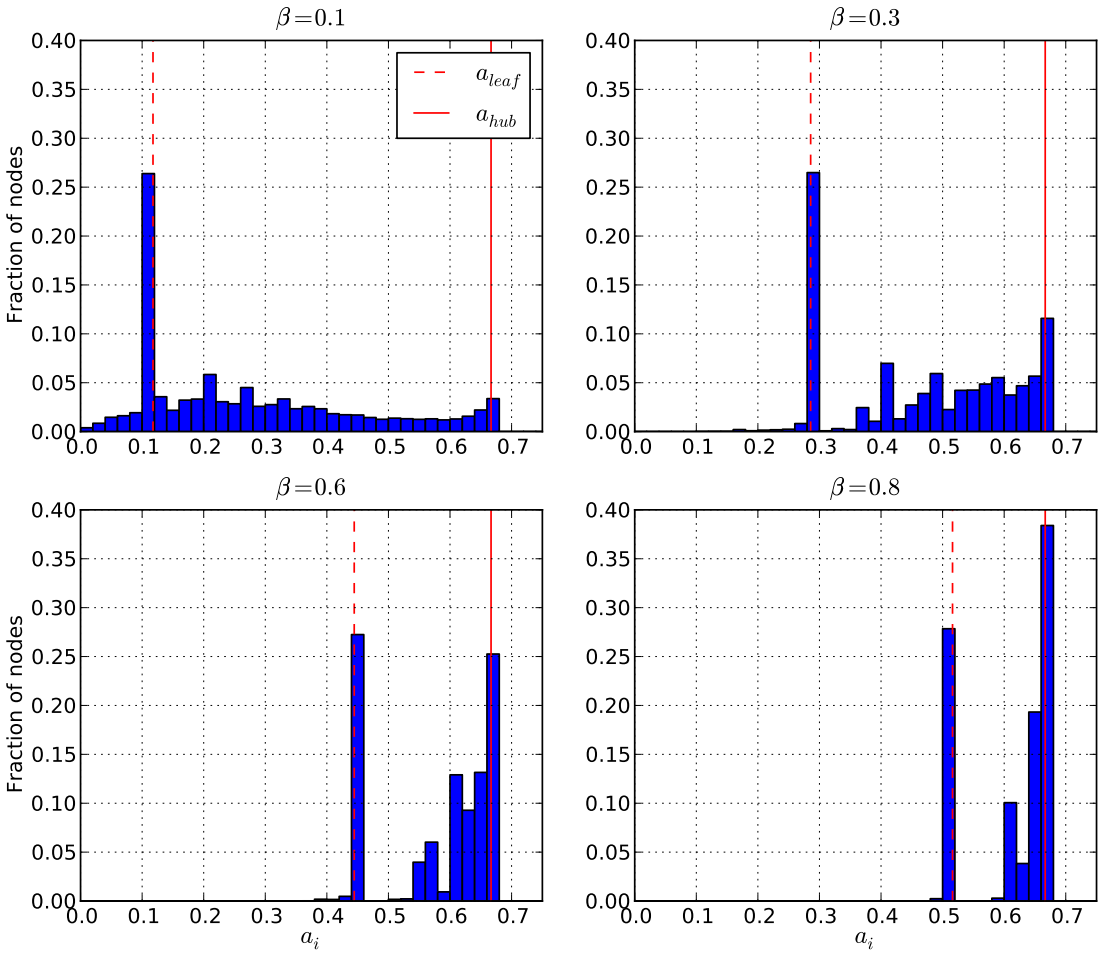

Figure: Histograms of infective nodes in the Enron e-mail network.

Real networks

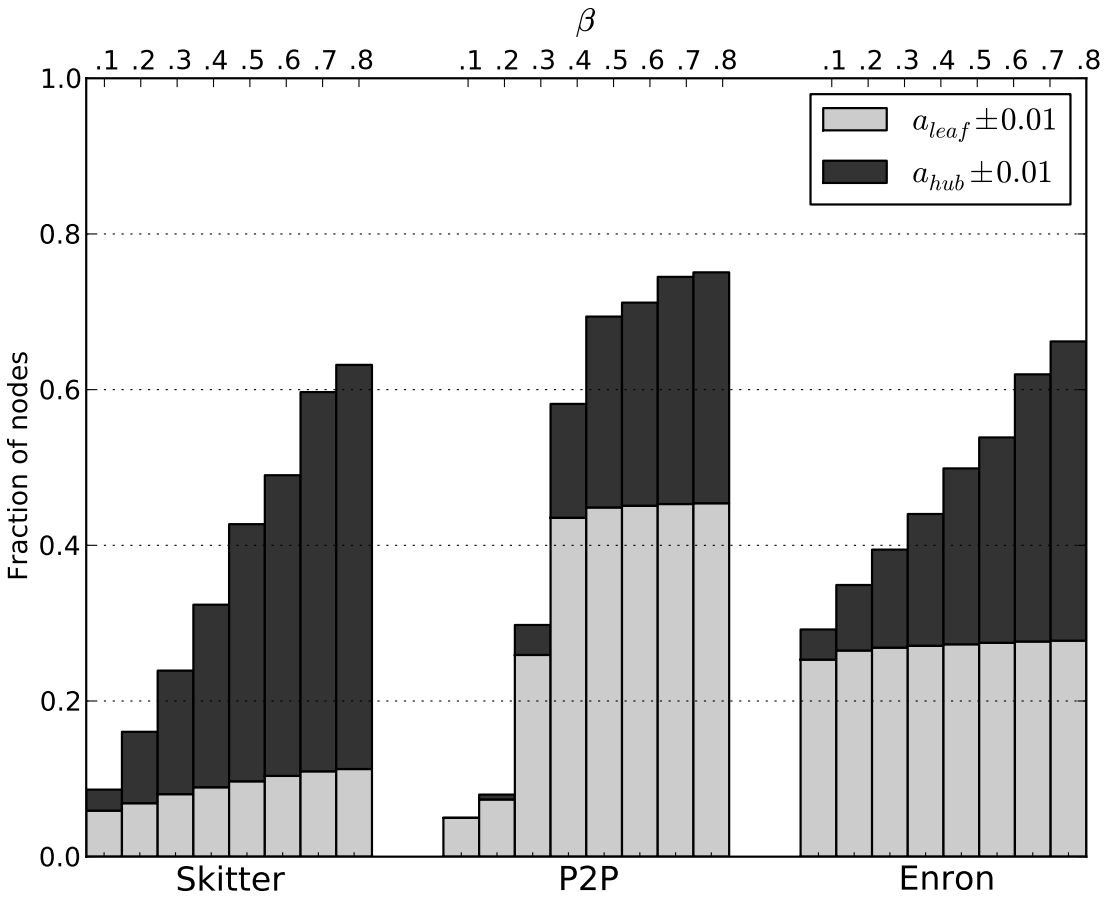

Figure: The fraction of nodes that behave as hubs (gray) and leaves (light gray) when β is varied, and $\delta = 0.5$, for three different networks.

12 different networks

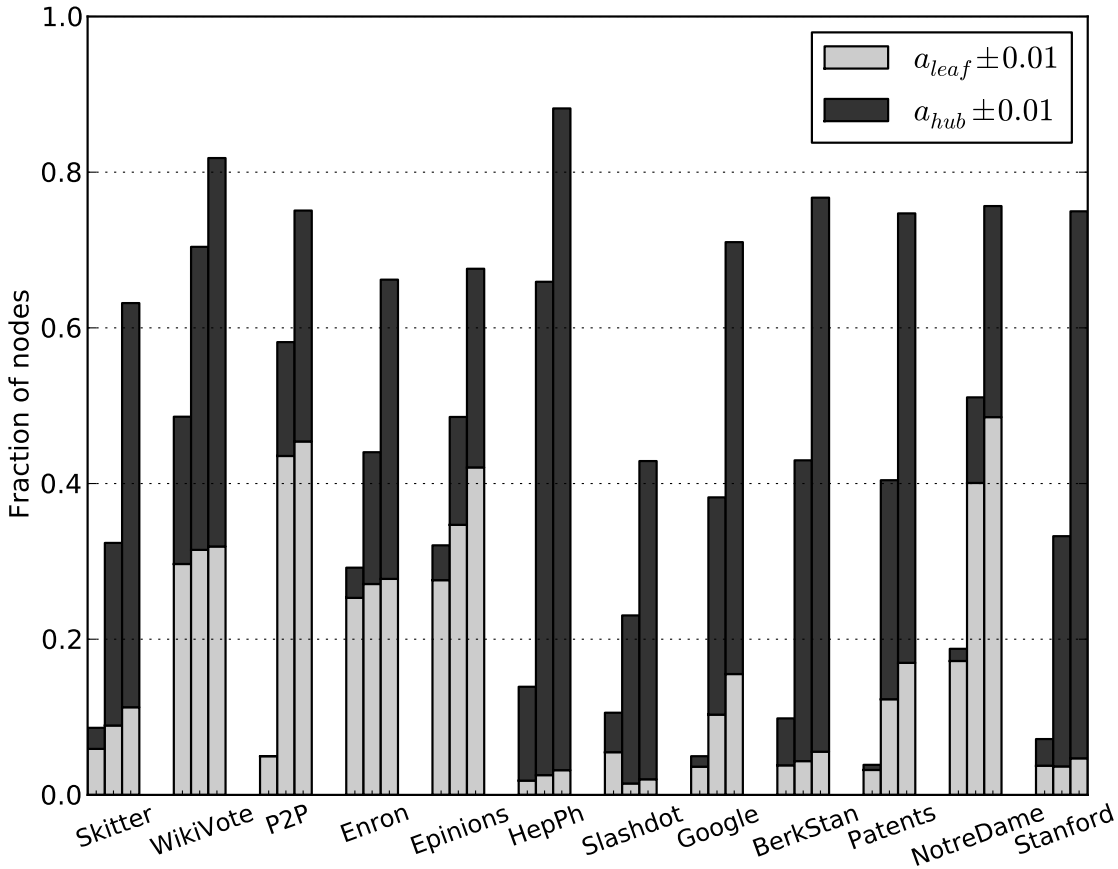

Figure: The fraction of nodes that behave as hubs (gray) and leaves (light gray) for $\beta = 0.1$, $\beta = 0.4$ and $\beta = 0.8$. $\delta = 0.5$.

Model of rumor spreading

$$p_i^I(t+1) = s_i^S(t) \sum_{j \in N(i)} \beta b_{ij} s_j^I(t) + (1-\delta) s_i^I(t)$$

- $N(i)$ – a set of all friends of a person i
- b_{ij} – probability that person i communicates with person j
- $\sum_j b_{ij} = 1$ – we assume that the person i certainly communicate with one person from the set $N(i)$
- β – probability of rumor transmission from an I node to an S node

Topology independent spreading processes

$$p_i^I(t+1) = \left[1 - p_i^I(t)\right] \sum_{j=1}^{N} \beta b_{ij} p_j^I(t) + (1-\delta) p_i^I(t)$$

It can be shown that in the limit case $p_i^I(\infty) = p^I$.

$$\begin{aligned} p^I &= (1-p^I)\beta p^I \sum_j b_{ij} + (1-\delta)p^I \\ &= (1-p^I)\beta p^I + (1-\delta)p^I \end{aligned}$$

The last equation has two solutions: $p^I = 0$ and $p^I = 1 - \frac{\delta}{\beta}$.

Topology independent spreading processes

- $\frac{\delta}{\beta} < 1$ – the solution $p^I = 1 - \frac{\delta}{\beta}$ is globally stable fixed point
- $\frac{\delta}{\beta} > 1$ – any infection in the network, will be eventually diminished, when $t \to \infty$
- Nodes status probabilities have an analytical solution in closed form: they are no longer topology dependent, and are functions only of the spreading process parameters β and δ for the SIS model

Summary

- Technological networks have quite advanced technological implementations, but our understanding of their flow mechanisms and long-term dynamics is far from complete
- Network science is a huge playground that can accommodate many research profiles: mathematicians, physicists, biologists, electrical and computer engineers, sociologists ...

References

- D. Trpevski, W. K. S. Tang, and L. Kocarev, Model for rumor spreading over networks, Physical Review E 81, 056102 (11 pages) 2010
- L. Kocarev and V. In, Network science: A new paradigm shift, IEEE Network, vol. 24 (6), pages: 6–9, 2010
- D. Smilkov and L. Kocarev, Analytically solvable processes on networks, submitted for publication, 2011
- I. Tomovski and L. Kocarev, Topology independent spreading processes, submitted for publication, 2011
- D. Trpevski, D. Smilkov, and L. Kocarev, On the number of infective nodes in epidemic models, submitted for publication, 2011

Thanks

Thanks to:

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