Identification of Influential Nodes in Social Networks

Maria G. Rossi, Fragkiskos D. Malliaros,
Michalis Vazirgiannis

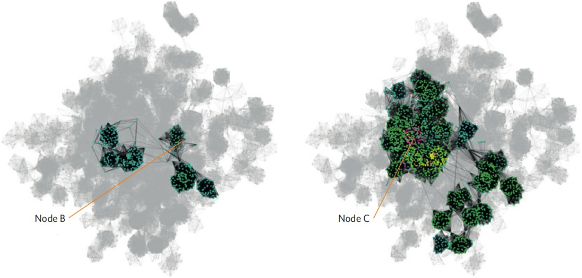
Department of Computer Science
École Polytechnique, France

DigiCosme Research Days - DataSense,
25 March 2015

Outline

Outline

Identifying influential spreaders
Goals



Find those nodes in the network that have a good influential power

Identifying influential spreaders

Goals

Goals

- Optimize the use of available resources
- Ensuring a more efficient spread of information
- In case of diseases hinder information spreading

Applications

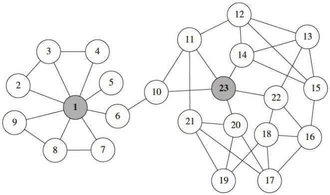
- epidemic control
- information diffusion
- viral marketing
- social movement
- idea propagation

Outline

Identifying influential spreaders

Related work

- degree centrality: straightforward metric to identify leaders in social networks
- high degree nodes may have low degree neighbors, hence hinder information spreading



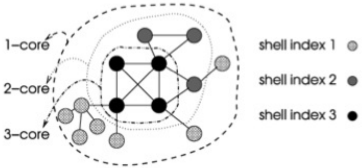
Chen, Duanbing, et al. "Identifying influential nodes in complex networks." Physica a: Statistical mechanics and its applications 391.4 (2012): 1777-1787.

Outline

k-core Decomposition

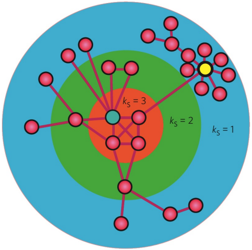
k-core Decomposition

- G = (V,E) undirected graph, V:number of nodes, E:number of edges
- Ck is the k-core subgraph of G in which all nodes have degree at least k
- C: set of nodes with the maximum core number k_{max}



k-core Decomposition

Most efficient spreaders are located within the *k*-core of the network



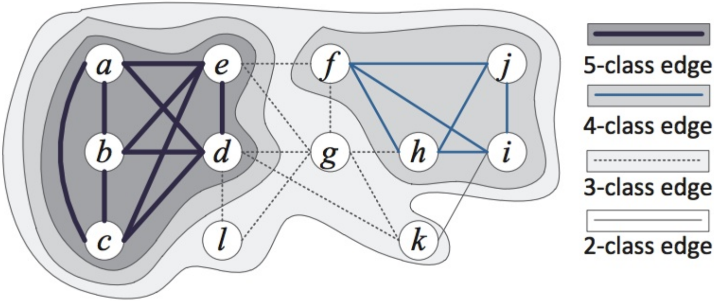
Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. Nature Physics, 6(11), 888-893.

Outline

K-Truss Decomposition

K-Truss Decomposition

T_K, $K \geq 2$: the K-truss subgraph of G , the largest subgraph where all edges belong to $K - 2$ **triangles**.



K-Truss Decomposition

Truss number $t_e = K$, Maximum node truss number T

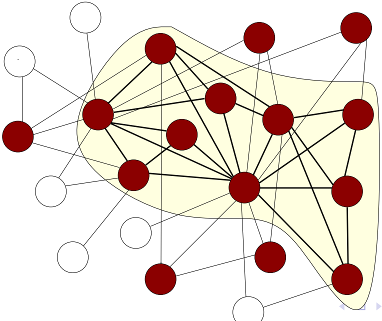
- $e \in E$ has truss number $t_e = K$ if it belongs to T_K but not to T_{K+1}
- $t_v, v \in V$ node's truss number as the maximum t_e of its adjacent edges
- T: the set of nodes with the maximum node truss number

Outline

k-core VS K-truss

k-core - K-truss relation

- Maximal k-core and K-truss subgraphs (i.e., maximum values for k, K) overlap
- K-truss is subgraph of k-core
- K-truss represents the *nucleus* of a k-core filtering out less important information.



k-core VS K-truss

T effect on spreading?

- How will spreading be affected if the epidemic starts from nodes belonging in set T (nodes of the max K-truss subgraph)?
- How will those nodes perform compared to the nodes in set C (nodes of the max k-core subgraph)?

Outline

SIR model

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$

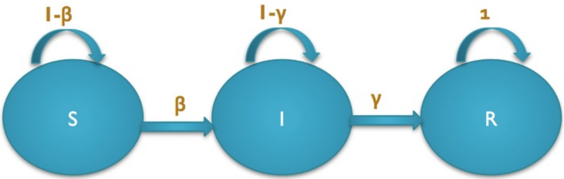
$$\frac{dR}{dt} = \gamma I$$

$S(t)$: number of Susceptible nodes
$I(t)$: number of Infected nodes
$R(t)$: number of Recovered nodes
β : infection rate
γ : recovery rate

Anderson, R. M.,& May, R. M. (1991). Infectious diseases of humans (Vol. 1). Oxford: Oxford university press.

SIR model

- Model for epidemics
- Individual node
- probabilistic transition among three states: Susceptible, Infected, Recovered (**SIR**)



Anderson, R. M.,& May, R. M. (1991). Infectious diseases of humans (Vol. 1). Oxford: Oxford university press.

Outline

Datasets

Dataset	$\lvert V \rvert$	$\lvert E \rvert$	k-core	K-truss	$\lvert C \rvert - \lvert T \rvert$	$\lvert T \rvert$	epid. thres.
Email-Enron	33.696	180.811	43	22	230	45	0.0084
epinions	75.877	405.739	67	33	425	61	0.0054
WikiVote	7.066	100.736	53	23	286	50	0.0072

Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J., & Faloutsos, C. (2008). Epidemic thresholds in real networks. ACM Transactions on Information and System Security (TISSEC), 10(4), 1.

Outline

Methodology

- Initiate the spreading process from a single node
- Repeat process 100 times for every seed node of each group:
  -the nodes belonging to the set T (**truss** method)
  -to those belonging to the set $C - T$ (**core** method)
  -those belonging to the set D that contains the highest degree nodes in the graph (**top degree** method)

Methodology

- Calculate the mean:
  i) number of nodes being infected at each step
  ii) the cumulative number of nodes
  iii) overall nodes' percentage infected at each step
- Spreading stops - store average and maximum number of steps
- Spreading parameters values: β - close to epidemic threshold $\tau = 1/\lambda_1$, $\gamma = 0.8$.

Outline

Results

Table 2: Average number of infected nodes per step of the SIR model using β close to the epidemic threshold of each graph and $\gamma = 0.8$. At the *Final step* column we show the total number of infected nodes at the end of the process (*Max step*).

		Time Step											
	Method	2	3	4	5	6	7	8	9	10	...	*Final step*	*Max step*
EMAIL-ENRON	**truss**	8.44	18.58	46.66	104.11	204.08	328.39	418.77	425.06	355.84	...	2,596.52	33
	core	4.78	12.82	31.97	73.77	152.55	264.36	367.28	403.98	364.13	...	2,465.60	37
	top degree	6.89	13.87	34.13	76.67	155.48	264.13	360.89	394.37	357.08	...	2,471.67	36
EPINIONS	**truss**	4.17	9.25	19.70	39.56	75.04	130.48	204.14	278.69	329.08	...	2,567.69	37
	core	3.45	7.18	14.72	29.11	55.27	98.11	158.56	226.17	280.03	...	2,325.37	43
	top degree	4.22	7.94	16.03	31.32	58.84	103.91	166.23	234.96	289.49	...	2,414.99	47
WIKI-VOTE	**truss**	2.92	4.37	6.92	10.43	15.27	21.63	28.73	35.93	42.46	...	560.66	52
	core	1.92	3.07	4.78	7.22	10.65	15.18	20.66	26.70	32.40	...	466.01	57
	top degree	2.43	3.53	5.46	8.17	12.05	17.04	23.05	29.49	35.55	...	502.88	62

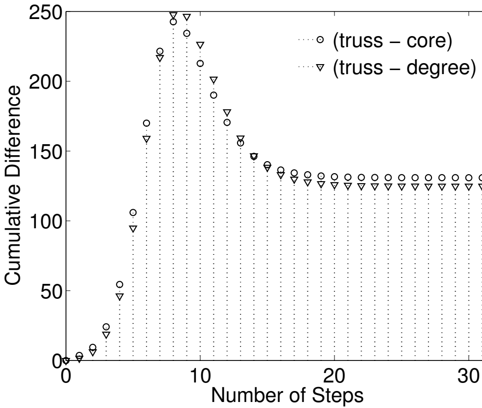
Maria-Evgenia G. Rossi, Fragkiskos D. Malliaros, and Michalis Vazirgiannis.
Spread it Good, Spread it Fast: Identification of Influential Nodes in Social Networks
International World Wide Web Conference (WWW), Florence, Italy, May 2015.

Results

Metrics

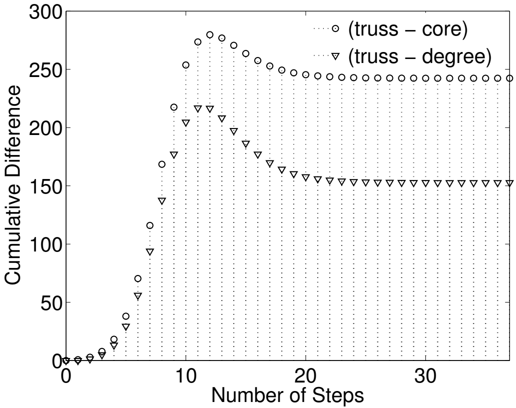
- $I_t^{\textbf{truss}}$: the number of infected nodes at step t by the **truss** method (similar for **core** and **top degree**).
- $\mathcal{D}_t^{\textbf{truss-core}} = \text{cumsum}_{z=1\ldots t}(I_z^{\textbf{truss}} - I_z^{\textbf{core}})$: the cumulative difference for the **truss** and **core** methods at step t as(similar for **truss** vs. **top degree**).

Results



(a) EMAIL-ENRON: $\beta = 0.01$

Results



(b) EPINIONS: $\beta = 0.007$

Outline

Benefits of K-truss vs. based k-core spreading

- During the first steps more nodes are infected: epidemy spreads faster
- Larger number of the infected nodes at the end of the process
- On average, spreading terminates earlier

Outline

Complexity issues

What about complexity?

- The k-core decomposition algorithm has linear complexity relative to the number of edges of the network , $O(n)$
- There exists a polynomial time algorithm for computing K-truss, $O(m^{1.5})$

Complexity issues

What about complexity?

- The K-truss algorithm has a higher time complexity than the k-core decomposition
- K-truss is a subgraph of k-core
- k-core computation complexity: linear and size(k-core) $<<$ size(Graph)
- Cohen et al. compute K-truss based on the k-core of the graph

Outline

Additional experiments

- **Multiple spreaders:**

  - Community detection
  - Choose as seed nodes those belonging to the k-core/K-truss subgraph of each community

- **Robustness of influential nodes under graph perturbations:**

  - Define noise model
  - Add noise to the graph
  - Examine how set of influential nodes are affected

Thank you!

Q&A