Spread it Good, Spread it Fast: Identification of Influential Nodes in Social Networks

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

Computer Science Laboratory, École Polytechnique, France

Introduction

Goal: find those nodes in the network that have a good influential power in order to:

- Optimize the use of available resources
- Ensure a more efficient spread of information
- Hinder information spreading (in case of diseases)

Related work:

- A straightforward metric to identify leaders in a social network would be the **degree centrality**
  ↪ But high degree nodes may have low degree neighbors, hence they eventually hinder information spreading
- It was shown that most efficient spreaders are located within the k-core of the network **[Kitsak et al., Nature Physics '10]**

Contributions:

- Refine the set of the most influential nodes, utilizing the properties of the K-truss decomposition – a triangle-based extension of k-core decomposition
- Locate nodes that perform faster and wider epidemic spreading

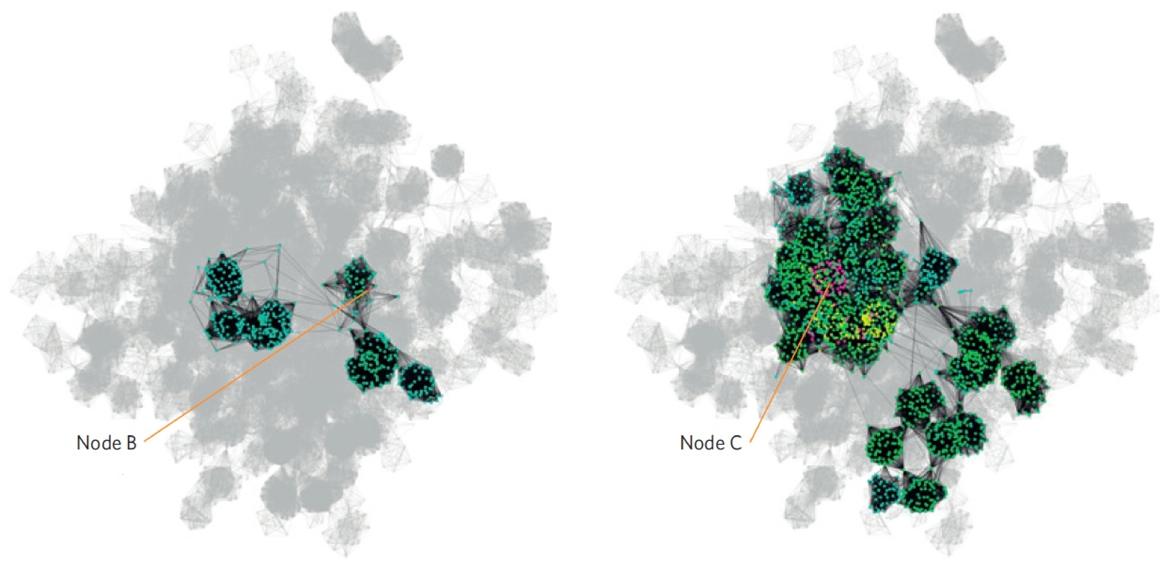


Figure : Influence of *Node B* having high degree and low k-core number versus influence of *Node C* having both high degree and k-core number **[Kitsak et al., Nature Physics '10]**

Preliminary Concepts and Definitions

DEFINITIONS: **k-core subgraph C_k, Core number c_v**:

- C_k is k-core subgraph of $G = (V, E)$ if it is a maximal connected subgraph in which all nodes have degree at least k
- Each node $v \in V$ has core number $c_v = k$, if it belongs to a k-core but not to a $(k+1)$-core

DEFINITIONS: **K-truss subgraph T_k, edge truss number t_e, node truss number t_v**:

- K-truss subgraph of $G = (V, E)$, denoted by T_K, $K \geq 2$, is defined as the largest subgraph where all edges belong to $K - 2$ triangles
- An edge $e \in E$ has truss number $t_e = K$ if it belongs to T_K but not to T_{K+1}
- The node's truss number t_v, $v \in V$ is the maximum t_e of its adjacent edges
- T denotes the set of nodes with the maximum node truss number

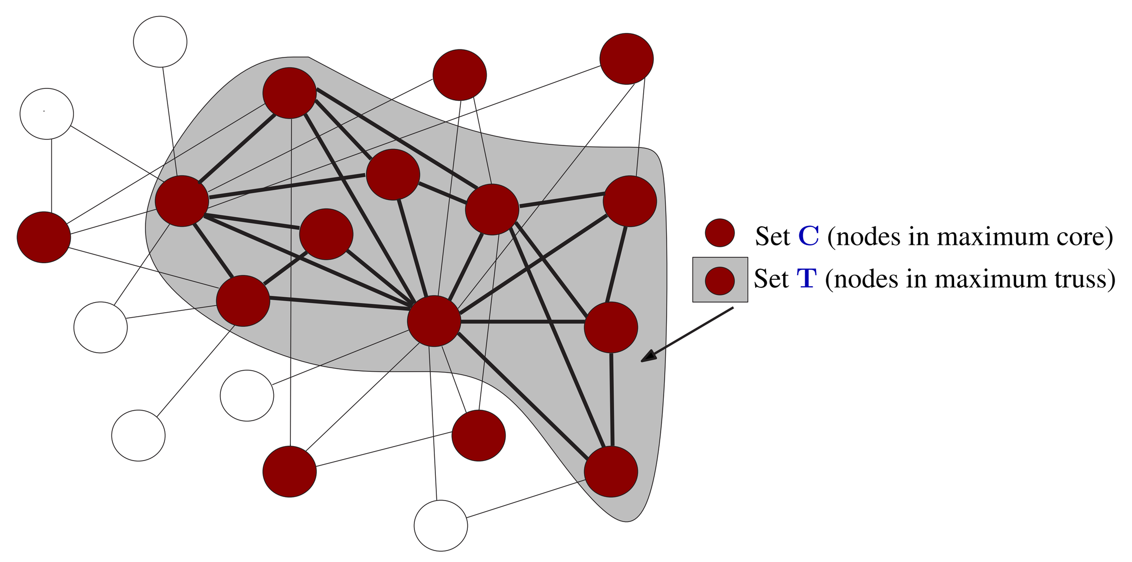


Figure : Maximal k-core and K-truss subgraphs (i.e., maximum values for k, K) overlap. We observe that K-truss – grey area – represents the *core* of a k-core – subgraph denoted by dark red nodes – that filters out less important information

Datasets

Network name	# Nodes	# Edges	k-core	K-truss	$\|C\| - \|T\|$	$\|T\|$	Epidemic threshold
EMAIL-ENRON	33,696	180,811	43	22	230	45	0.0084
EPINIONS	75,877	405,739	67	33	425	61	0.0054
WIKI-VOTE	7,066	100,736	53	23	286	50	0.0072

Experimental Evaluation (I)

		Time Step					
	Method	2	6	10	*Final step*	σ	*Max step*
EMAIL-ENRON	**truss**	8.44	204.08	355.84	2,596.52	136.7	33
	core	4.78	152.55	364.13	2,465.60	199.6	37
	top degree	6.89	155.48	357.08	2,471.67	354.8	36
EPINIONS	**truss**	4.17	75.04	329.08	2,567.69	227.8	37
	core	3.45	55.27	280.03	2,325.37	327.2	43
	top degree	4.22	58.84	289.49	2,414.99	331.7	47
WIKI-VOTE	**truss**	2.92	15.27	42.46	560.66	114.9	52
	core	1.92	10.65	32.40	466.01	104.5	57
	top degree	2.43	12.05	35.55	502.88	104.5	62

Table : Average number of infected nodes for some steps of the SIR model, using β close to the epidemic threshold of each graph and $\gamma = 0.8$

- The **truss** method achieves higher infection rate during the first steps
- The total number of infected nodes at the end of the process is larger, while the fade out occurs earlier

Acknowledgements

DigiCosme Ph.D. Fellowship (Maria-Evgenia G. Rossi)
Google Europe Fellowship in Graph Mining (Fragkiskos D. Malliaros)

Proposed Methodology

- Aim to identify those **single spreaders** in a network that will achieve an efficient spreading of information
- Argue that those nodes are located in the node set T of the graph, produced by the K-truss decomposition

DEFINITIONS: node sets C, T and D:

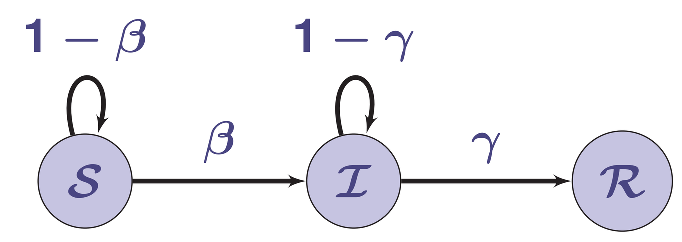
- C is the set of nodes with the maximum core number
- T is the set of nodes with the maximum node truss number
- D is the set of highest degree nodes of the network

DEFINITIONS: methods **truss**, **core** and **top degree**:

- **truss** method: nodes belonging to the set T
- **core** method: nodes belonging to the set $C - T$
- **top degree** method: nodes belonging to the set D

How to simulate the spreading process?

- We apply the **S**usceptible-**I**nfected-**R**ecovered (SIR) model **[Easley & Kleinberg, Cambridge University Press '10]**
  - Set one node to be infected (single spreader), as chosen by different methods
  - Infected nodes can infect their susceptible neighbors with probability β
  - A node that has been previously infected can recover from the disease with a probability γ



Experimental Evaluation (II)

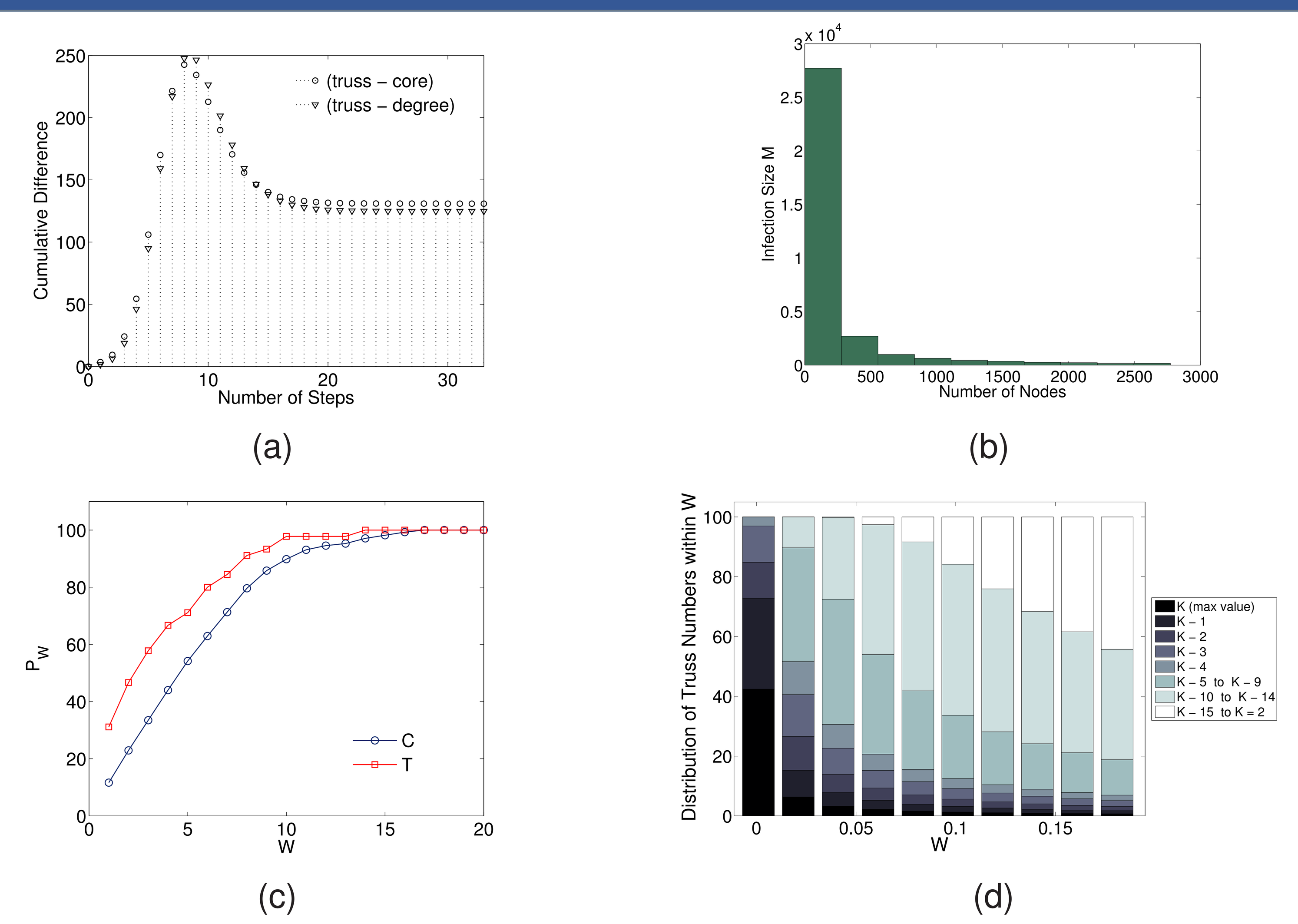


Figure : Results on the EMAIL-ENRON dataset. (a) Cumulative differences of infected nodes per step, (b) Spreading distribution of nodes, (c) Distribution of top-truss $\mathcal{P}_W^T$ and top-core $\mathcal{P}_W^C$ nodes within window W, (d) Distribution of node's truss number within window W

- Rank nodes according to the spreading M that they achieve → For small values of window size W, the number of top-truss nodes is always higher than the number of top-core nodes
- For small window sizes (i.e., close to the optimal spreading), the groups of nodes having high truss number conquer the set

{rossi, fmalliaros, mvazirg}@lix.polytechnique.fr