INFLUENTIAL NODE DISCOVERY IN SOCIAL NETWORKS

► *Azim Ahmadzadeh* | ► *Sajjad Hassanpour* | ■ *Dr.Tomasz Brengos*

Contribution

For an experimental study, we develope a desktop application which runs experiments for any given network to assist us with evaluating the effectiveness of the **Flow Authorities Model** which is proposed to discover a highly reliable solution for the **Campaign problem**. We would like to find an answer for this question that whether or not this solution is suitable for protecting a network against the spread of a virus as well.

Campaign Problem

Determine the set S of K data points at which the release of information bits I would maximize the expected number of nodes over which I is assimilated.[1]

The Question

Since S is very carefully selected to maximize the spread of the information, it seems to be an intuitively and reasonably good collection of nodes to be chosen when it comes to protecting the network against attacks or viruses. In other words:

> Does protection of S result in the best protection of the network, for a given K?

Flow Authorities Model

This model [1] proposes a method that is claimed to discover a highly reliable solution for the Campaign Problem. It is based on the *Steady State Probability* of each node having the information. Using the *Markov property* of the system, the expected number of nodes exposed to information can be derived from the system of equations below:

$$\begin{cases} 1-\pi(i) = \prod_{l\in N(i)}(1-\pi(l)\cdot p_{li}) \\ \pi(i) = 1 \qquad \forall i \in S \end{cases}$$

where p_{li} is the transition probability from a neighboring node l to i, and $\pi(i)$ is the probability that the node i has the information. In addition, $\pi(i) = 1$ means that i initially has the information.

Virus Spread Model

Our Virus Spread Model follows the *Linear Threshold* pattern. The probability that j becomes infected is proportionate to the probability that one of its neighbors i is already infected multiplied by the transition probability p_{ij}.

► References

[1] Charu C. Aggarwal, Arijit Khan and Xifeng Yan *On Flow Authority Discovery in Social Networks*, 2011

[2] Aristides Gionis, Evimaria Terzi, and Panayiotis Tsaparas *Opinion maximization in social networks*, 2013

[3] Allan Borodin, Yuval Filmus, and Joel Oren *Threshold Models for Competitive Influence in Social Networks*, 2010

[4] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze *Introduction to Information Retrieval*, 2008

[5] Lavon B. Page, Sandra O. Paur *The NCSU Finite Math Book*, 1989

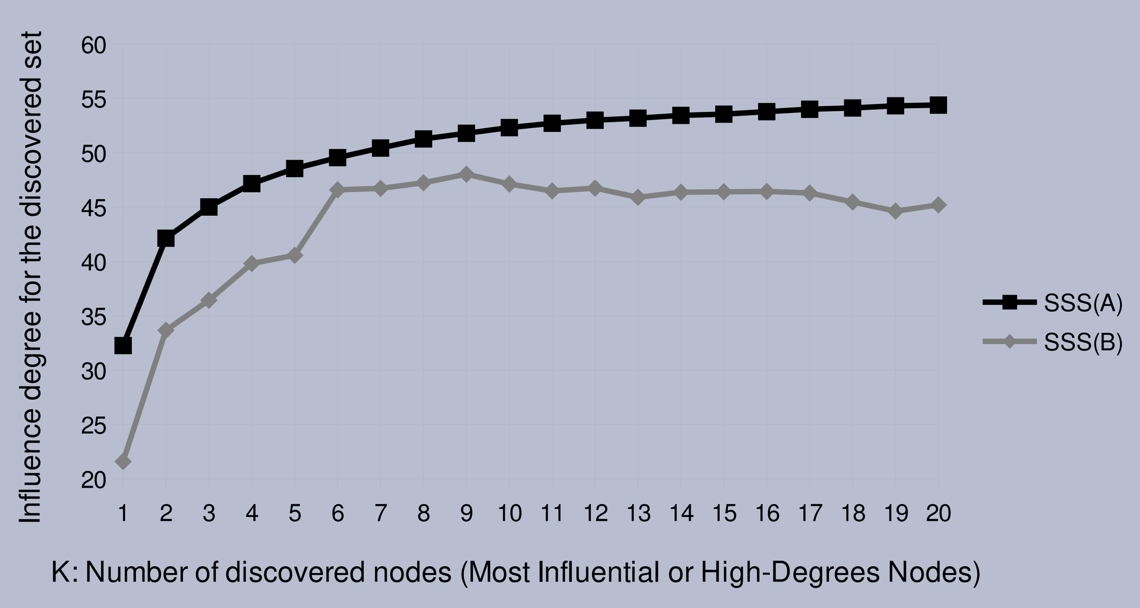
Designed Experiment

Our experiment relies on our virus spread model. For any given network, two copies are created: A and B. In A only the K Most Influential nodes are vaccinated and in B only the nodes with the K highest number of outgoing degrees. We start the experiment by injecting the virus to both networks through a unique set of nodes and when the spread ceased we compare the infection levels.

Expected Results

We expect to observe a more significant impact on A than B. So, the degree of the spread in the network A where the claimed most influential nodes are vaccinated and unable to help the spread, is expected to drop to a lower level than what the network B experiences.

General Behavior



The bahavior of the *Most Influential* set (■) is slightly different than the *High-Degrees* set (♦), in terms of their overall influence. For any K, the Most Influential set has a greater impact on the network. Although, they both has a logarithmic behavior, it is only the Most Influential set that by increasing K, always exerts more influence on the network. In case of the High-Degrees set, increasing K does not always result in a more influential set.

Experiments Assumptions

In our experiment, several assumptions are held to assure unbiasedness. Given that:

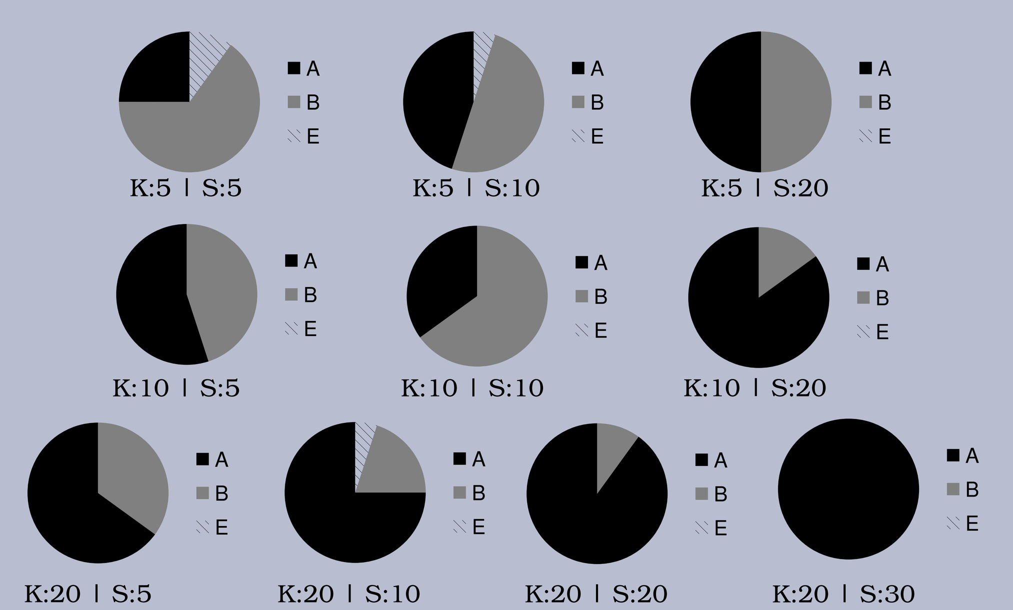
- M_{Inf} : Most Influential set
- H_{Deg} : High-Degrees set
- I_{Ini} : Initially Infected set
- I_{Eve} : Eventually Infected set
- Vac : Vaccinateed set

then the assumptions are:

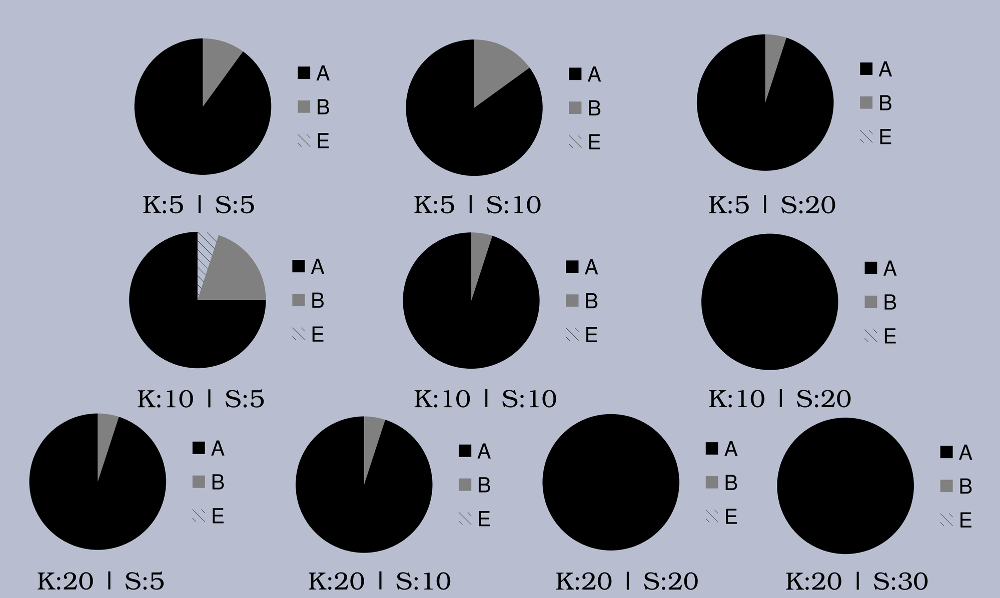
1. $(M_{Inf} \subseteq V) \wedge (H_{Deg} \subseteq V)$
2. $M_{Inf} \cap H_{Deg} = \emptyset$
3. $|M_{Inf}| = |H_{Deg}| = K$
4. $I_{Ini} \subseteq I_{Eve}$
5. $Vac \cap I_{Ini} = \emptyset$
6. $Vac \cap I_{Eve} = \emptyset$
7. $(Vac = M_{Inf}) \veebar (Vac = H_{deg})$

Results

We run our experiments on a network of the US airports with 332 nodes which are connected with weighted edges.
After 200 experiments, despite our expectation, not only we do not see the network B to *always* experience a higher infection degree but also it seems that as K increases, A gets more vulnerable against the virus. It means that in a society where the Most Influential nodes are unable to play their critical role in spreading information bits, yet the virus spreads in a greater scale than the society in which the High-Degrees nodes are unable to participate.



Relative frequency of the experiment (A or B) with higher infection rate before the skewness correction



Relative frequency of the experiment (A or B) with higher infection rate after the skewness correction

The Flow Authorities Model discovers the most influential nodes that indeed form a highly reliable set of nodes to be employed if the goal is to spread the information bits. But now we can confidently state that this set of nodes is absolutely not the right selection to be protected if the goal is to minimize the spread.

> The most influential nodes discovered by the Flow Authorities Model must not be used in any practices that aim at protecting a network from the spread of the information bits.

■ About

Bachelor Thesis | *Computer Science* | May 2016
Warsaw University of Technology | *Mathematics and Information Science Faculty*