Identification of Influential Nodes in Social Networks

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

Department of Computer Science
Ecole Polytechnique, France

DigiCosme Research Days - DataSense,
25 March 2015
Outline

1. Identifying influential spreaders
   - Goals
   - Related work

2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss

3. The epidemic model
   - The SIR model

4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues

5. Ongoing work
   - Additional experiments
Outline

1. Identifying influential spreaders
   - Goals
   - Related work
2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss
3. The epidemic model
   - The SIR model
4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues
5. Ongoing work
   - Additional experiments
Identifying influential spreaders

Goals

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

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

1. Identifying influential spreaders
   - Goals
   - Related work

2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss

3. The epidemic model
   - The SIR model

4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues

5. Ongoing work
   - Additional experiments
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

Outline

1. Identifying influential spreaders
   - Goals
   - Related work

2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss

3. The epidemic model
   - The SIR model

4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues

5. Ongoing work
   - Additional experiments
k-core Decomposition

- G = (V, E) undirected graph, V: number of nodes, E: number of edges
- C_k 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.

Outline

1. Identifying influential spreaders
   - Goals
   - Related work

2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss

3. The epidemic model
   - The SIR model

4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues

5. Ongoing work
   - Additional experiments
**K-Truss Decomposition**

A subgraph T_K, K ≥ 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
**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

1. Identifying influential spreaders
   - Goals
   - Related work

2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss

3. The epidemic model
   - The SIR model

4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues

5. Ongoing work
   - Additional experiments
SIR model

\[
\begin{align*}
\frac{dS}{dt} &= -\frac{\beta SI}{N} \\
\frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I \\
\frac{dR}{dt} &= \gamma I
\end{align*}
\]

- \(S(t)\) : number of Susceptible nodes
- \(I(t)\) : number of Infected nodes
- \(R(t)\) : number of Recovered nodes
- \(\beta\) : infection rate
- \(\gamma\) : recovery rate

**SIR model**

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

Outline

1. Identifying influential spreaders
   - Goals
   - Related work

2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss

3. The epidemic model
   - The SIR model

4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues

5. Ongoing work
   - Additional experiments
## Datasets

| Dataset       | |V| | E | k-core | K-truss | |C| – | T | |T| | 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   |

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: \( \beta \) - close to epidemic threshold \( \tau = 1/\lambda_1 \), \( \gamma = 0.8 \).
Outline

1. Identifying influential spreaders
   - Goals
   - Related work
2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss
3. The epidemic model
   - The SIR model
4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues
5. Ongoing work
   - Additional experiments
## Results

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

Results

Metrics

- $I_t^{\text{truss}}$: the number of infected nodes at step $t$ by the \textbf{truss} method (similar for \textbf{core} and \textbf{top degree}).

- $D_t^{\text{truss-core}} = \text{cumsum}_{z=1}^{t}(I_z^{\text{truss}} - I_z^{\text{core}})$: the cumulative difference for the \textbf{truss} and \textbf{core} methods at step $t$ as (similar for \textbf{truss} vs. \textbf{top degree}).
Results

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

(b) EPINIONS: $\beta = 0.007$
Outline

1. Identifying influential spreaders
   - Goals
   - Related work

2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss

3. The epidemic model
   - The SIR model

4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues

5. Ongoing work
   - Additional experiments
Benefits of $K$-truss vs. based $k$-core spreading

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

1. Identifying influential spreaders
   - Goals
   - Related work

2. Graph Degeneracy and Influential Spreaders
   - k-core Decomposition
   - K-Truss Decomposition
   - k-core VS K-truss

3. The epidemic model
   - The SIR model

4. Experiments
   - Datasets used
   - Methodology
   - Results
   - Benefits
   - Complexity issues

5. Ongoing work
   - Additional experiments
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