Large Social Networks can be Targeted for Viral Marketing with Small Seed Sets

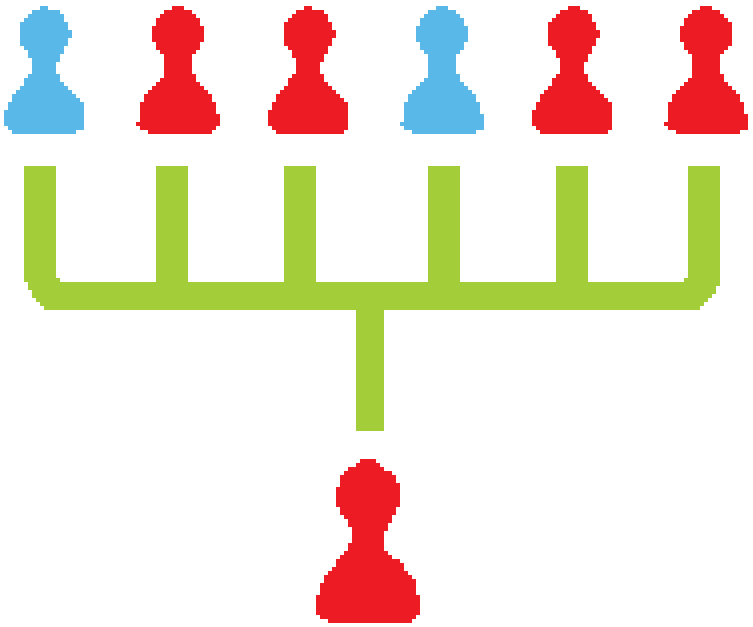
ASONAM 2012

Paulo Shakarian and Damon Paulo
Dept. Electrical Engineering and Computer Science and
Network Science Center
U.S. Military Academy
West Point, NY

Network Science Center
West Point

Tipping

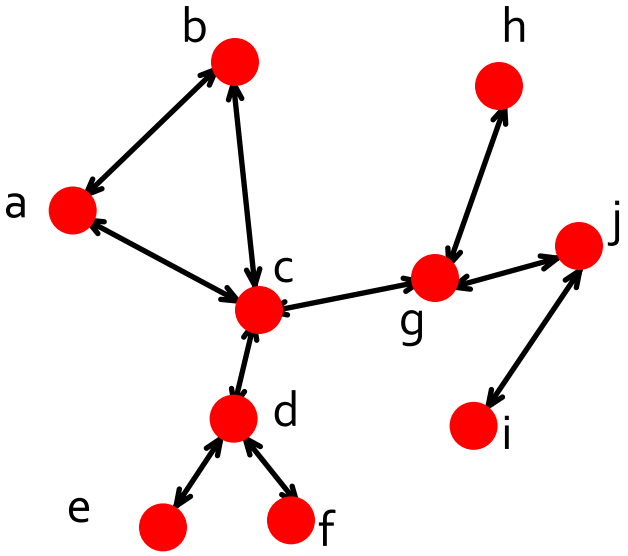
- The "tipping point" is often referred to in popular literature.
  - First introduced in Schelling '78 and Granovetter '78
- In layman's terms:
  - Every member of the population has a threshold ***k***
  - If ***k*** number of your friends adopt a behavior – then you will too.



Network Science Center
West Point

Tipping

- The idea of "tipping" can be extended to social networks to describe a diffusion process.
- We will often say a node is "activated" if it has adopted a certain behavior.



Network Science Center
West Point

The Tipping Point in the Early 2000's

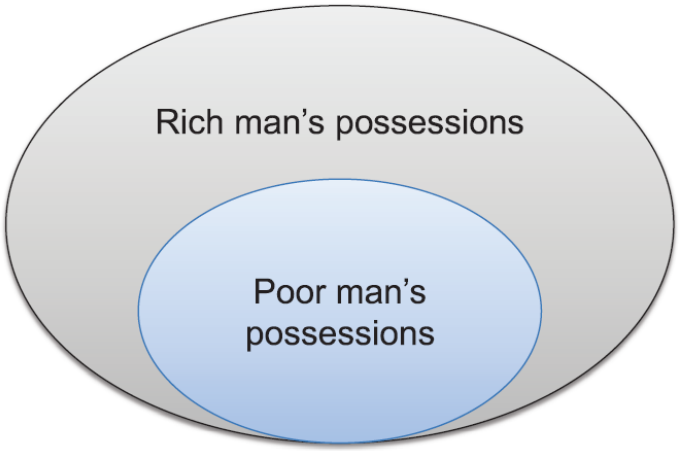
- In 2003, Kempe, Kleinberg, and Tardos introduce a framework for dealing with the tipping problem.

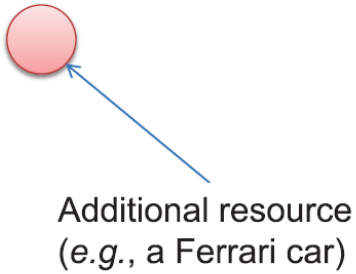
- They state that the parameter ***k*** can often not be determined, and hence assigned based on some probability distribution.

The Tipping in the KKT Model

- They show that the cascade initiated by tipping is ***submodular*** when ***k*** is assigned probabilistically.

- Intuitively, a submodular function has diminishing returns





The Tipping in the KKT Model

- The submodularity of the tipping process in the non-deterministic KKT model was leveraged in the proof of an approximation guarantee for a greedy algorithm to find the most influential nodes (w.r.t expected value).

- However, their approach was not scalable to huge social networks as their approach relied on simulation runs.

Network Science Center
West Point

Recent Work on Influence

- Recent studies of very large social networks has suggests that we may be able to determine ***k*** based on the attributes of a given node.

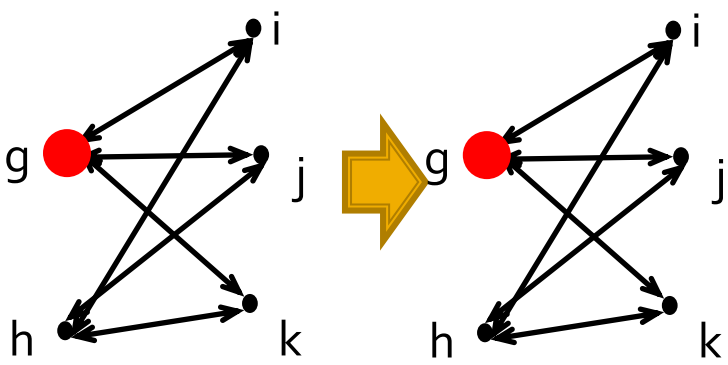
- For instance, Aral and Walker (2012) :
  - People over 31 are less susceptible to influence than younger people
  - Men are more susceptible to influence than women
  - Married people are less susceptible to influence than people engaged to be married

Network Science Center
West Point

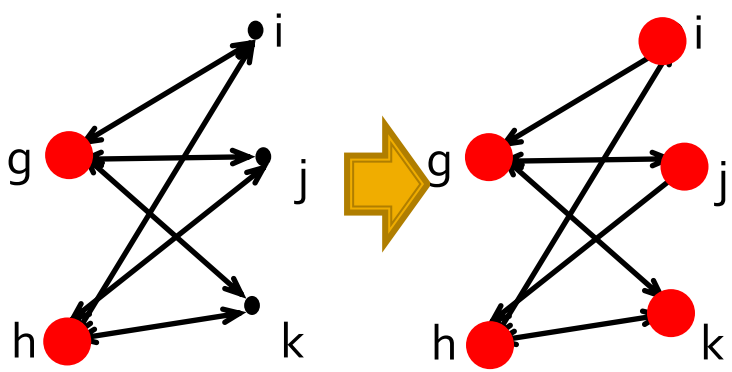
Recent Work on Influence

However, submodularity does not hold if the threshold is deterministic

k=2 for all nodes



Incremental increase of adding node g to the empty set: 1



Incremental increase of adding node g to set {h} : 4

Network Science Center West Point

Research Questions

- If we know the ***k*** value for each node, can we find a (near-minimal) set of nodes that causes universal adoption under the tipping model?
- Can we find such a set in very large networks?

Outline

- Preliminaries
- The Pruning Heuristic
- Experimental Evaluation
- Concluding Remarks

Preliminaries

We assume a directed network and a threshold function that provides the percentage of "active" neighbors required to tip that node.

$G = (V, E)$

$For\ all\ v_i,\ \eta_i^{in}$ is the set of incoming neighbors

$\theta : V \rightarrow (0,1]$

$For\ all\ v_i \in V,\ k_i = \lceil \theta(v_i)\ d_i^{in} \rceil$

Network Science Center
West Point

Activation Function

- Maps subsets of vertices (nodes activated at the current time step) to subsets of vertices (nodes activated at the next time step)

$$A_\theta(V') = V' \cup \{v_i \in V \ s.t. \ |\eta_i^{in} \cap V'| \geq k_i\}$$

- Defined for multiple iteration below:

$$A_\theta^i(V') = \begin{cases} A_\theta(V') & \text{if } i = 1 \\ A_\theta(A_\theta^{i-1}(V')) & \text{otherwise} \end{cases}$$

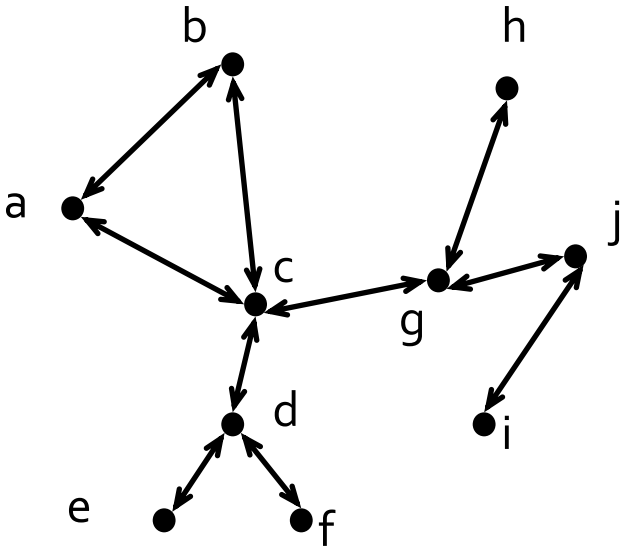
Network Science Center
West Point

The Activation of Nodes

Consider the social network to the left.

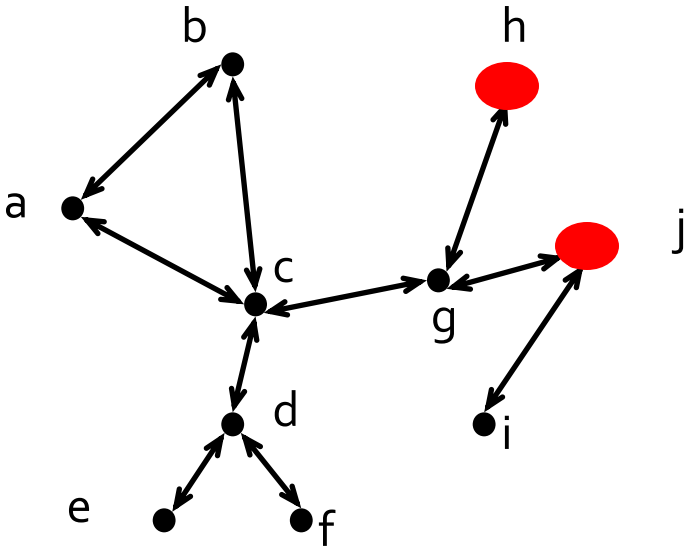
Suppose we have a majority threshold.

$For\ all\ v_i \in V,\ \theta(v_i)=0.5$



The Activation of Nodes

Initially individuals h and j are activated.

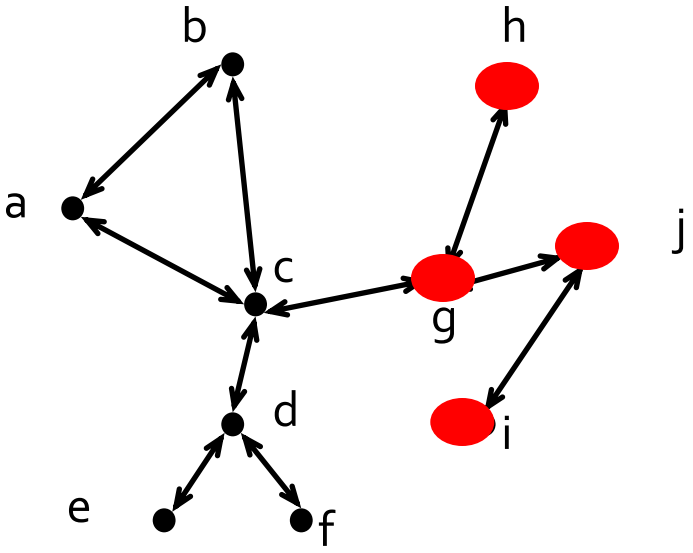


The Activation of Nodes

Initially villages **h** and **j** are activated.

This leads to **g** and **i** also becoming activated in the next time step.

But no others.



Problem

- Clearly, the tipping process must complete in less than $|V|$ time steps
- For $j > |V|$, we define the following:

$$\mathbf{\Gamma}_\theta(V') = A_\theta^j(V')$$

- The *MIN-SEED* problem is to find a set $V' \subseteq V$ of minimal size that causes $\Gamma_\theta(V') = V$.

- Previously, this problem has been shown to be **NP-Complete.** (Dreyer & Roberts, '09)

Network Science Center
West Point

The Minimum Seed Problem

Our new pruning heuristic guarantees that the result will lead to adoption of the behavior by the entire population, but does not guarantee minimal size.

However, experimentally, we find that the set returned is often several orders of magnitude smaller than the population.

The Pruning Heuristic for the Tipping Model

Algorithm 1 TIP_DECOMP

Require: Threshold function, θ and directed social network $G = (V, E)$

Ensure: V'

For each vertex v_i, compute k_i.
For each vertex v_i, $dist_i = d_i^{in} - k_i$.
FLAG = TRUE.
while FLAG **do**
  Let v_i be the element of v where $dist_i$ is minimal.
  if $dist_i = \infty$ **then**
    FLAG = FALSE.
  else
    Remove v_i from G and for each v_j in η_i^{out}, if $dist_j > 0$, set $dist_j = dist_j - 1$. Otherwise set $dist_j = \infty$.
  end if
end while
return All nodes left in G.

Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

- **<u>THEOREM:</u>** Given set V' returned by TIP_DECOMP, $\Gamma_\theta(V')=V$.

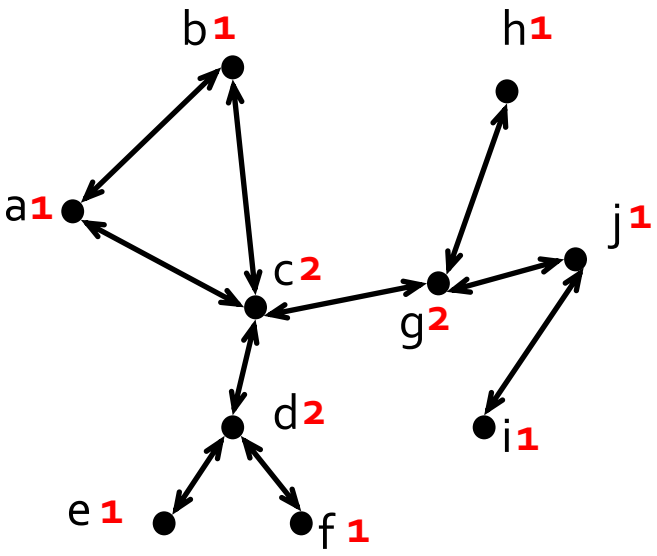
- **<u>PROPOSITION:</u>** On a directed network of N nodes and M edges, TIP_DECOMP runs in $O(M \lg N)$ time.

Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

Let us consider the following network.

First, our algorithm identifies the threshold for each node to be tipped. If we assume at least 50% of each neighbor is required, then this number is shown in red.

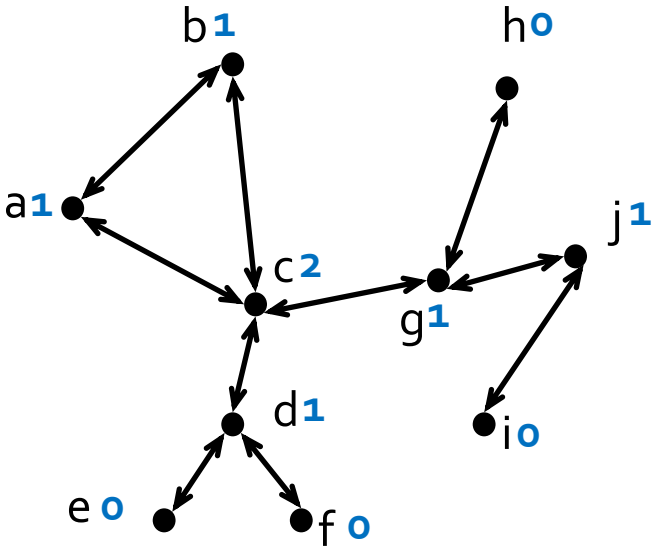


Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

Next, the distance to being tipped is computed. This is defined as the node's current degree minus the threshold.

It is shown for each node in blue



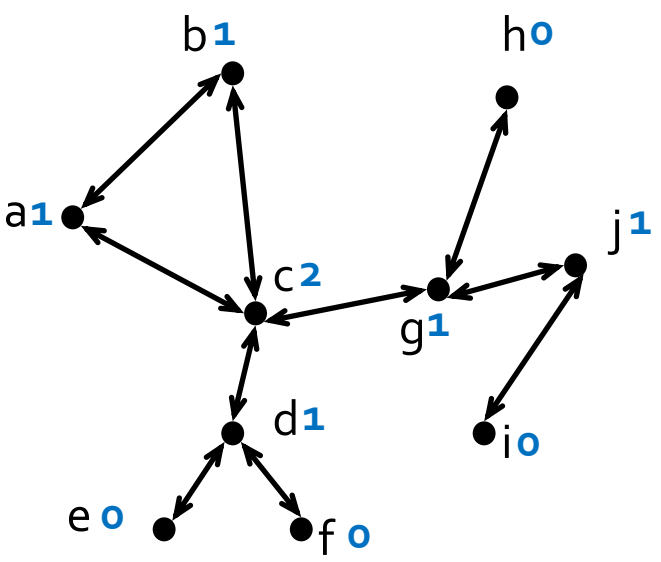
Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

The algorithm then proceeds as follows:

At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

Then update the distances of its neighbors.



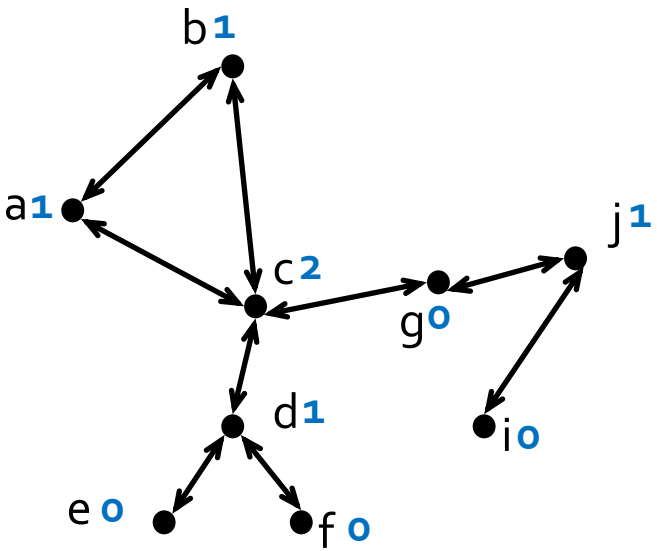
Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

The algorithm then proceeds as follows:

At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

Then update the distances of its neighbors.



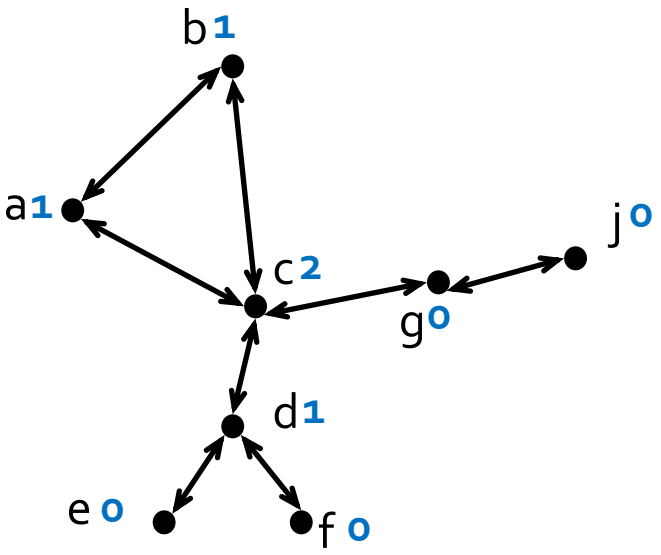
Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

The algorithm then proceeds as follows:

At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

Then update the distances of its neighbors.



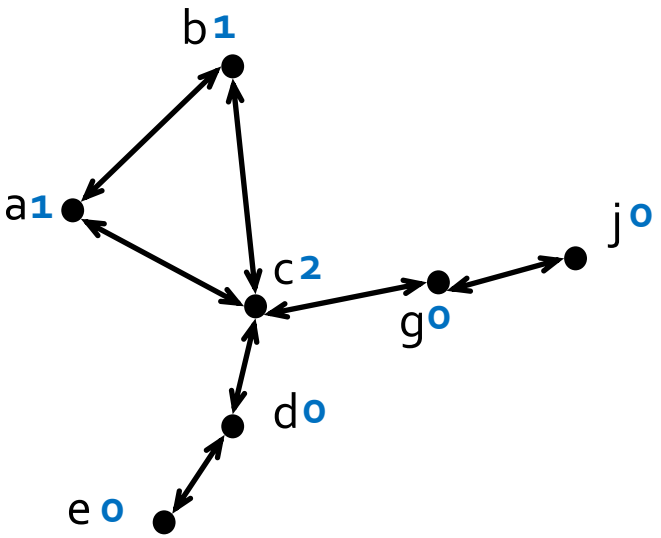
Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

The algorithm then proceeds as follows:

At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

Then update the distances of its neighbors.



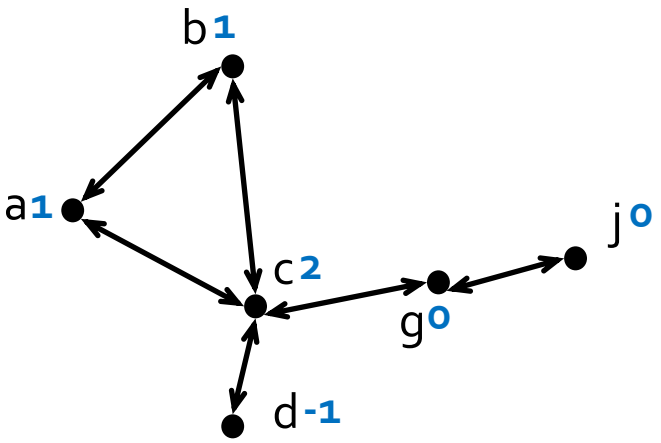
Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

The algorithm then proceeds as follows:

At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

Then update the distances of its neighbors.



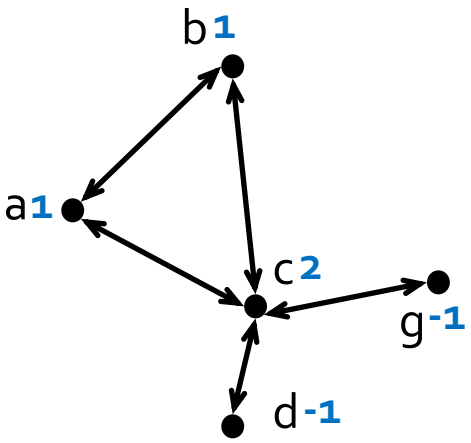
Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

The algorithm then proceeds as follows:

At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

Then update the distances of its neighbors.



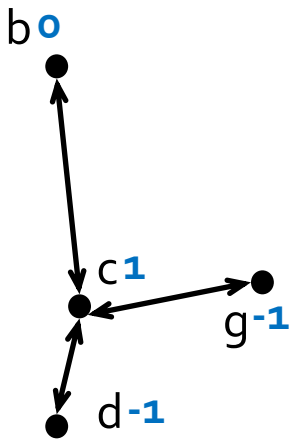
Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

The algorithm then proceeds as follows:

At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

Then update the distances of its neighbors.



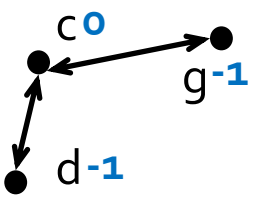
Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

The algorithm then proceeds as follows:

At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

Then update the distances of its neighbors.



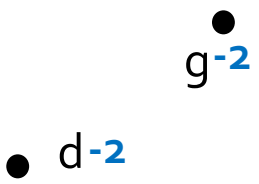
Network Science Center
West Point

The Pruning Heuristic for the Tipping Model

The algorithm then proceeds as follows:

At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

Then update the distances of its neighbors.



Solution = { d, g }

Network Science Center
West Point

Experimental Evaluation

- TIP_DECOMP was implemented in about 200 lines of Python code using the NetworkX library.
- Experiments performed on 31 real-world social networks

Category A Networks

CATEGORY A

Name	# Nodes	# Edges	Avg. Degree	Source	Type
BlogCatalog1	88,784	4,186,390	23.58	ASU	SocMedia
BlogCatalog2	97,884	3,337,294	17.05	ASU	SocMedia
BlogCatalog3	10,312	667,966	32.39	ASU	SocMedia
Buzznet	101,163	5,526,132	27.31	ASU	SocMedia
Douban	154,908	654,324	2.11	ASU	SocMedia
Flickr	80,513	11,799,764	73.28	ASU	SocMedia
Flixster	2,523,386	15,837,602	3.14	ASU	SocMedia
FourSquare	639,014	6,429,972	5.03	ASU	SocMedia
Frienster	5,689,498	28,135,774	2.47	ASU	SocMedia
Last.Fm	1,191,812	9,038,680	3.79	ASU	SocMedia
LiveJournal	2,238,731	25,632,368	5.72	ASU	SocMedia
Livemocha	104,103	4,386,166	21.07	ASU	SocMedia
WikiTalk	2,394,385	9,319,130	1.95	SNAP	SocMedia

Category B Networks

CATEGORY B

Name	# Nodes	# Edges	Avg. Degree	Source	Type
Delicious	536,408	2,732,272	2.55	ASU	SocMedia
Digg	771,231	11,814,826	7.66	ASU	SocMedia
EU E-Mail	265,214	728,962	1.37	SNAP	E-Mail
Hyves	1,402,673	5,554,838	1.98	ASU	SocMedia
Yelp	487,401	4,686,962	4.81	ASU	SocMedia

Category C Networks

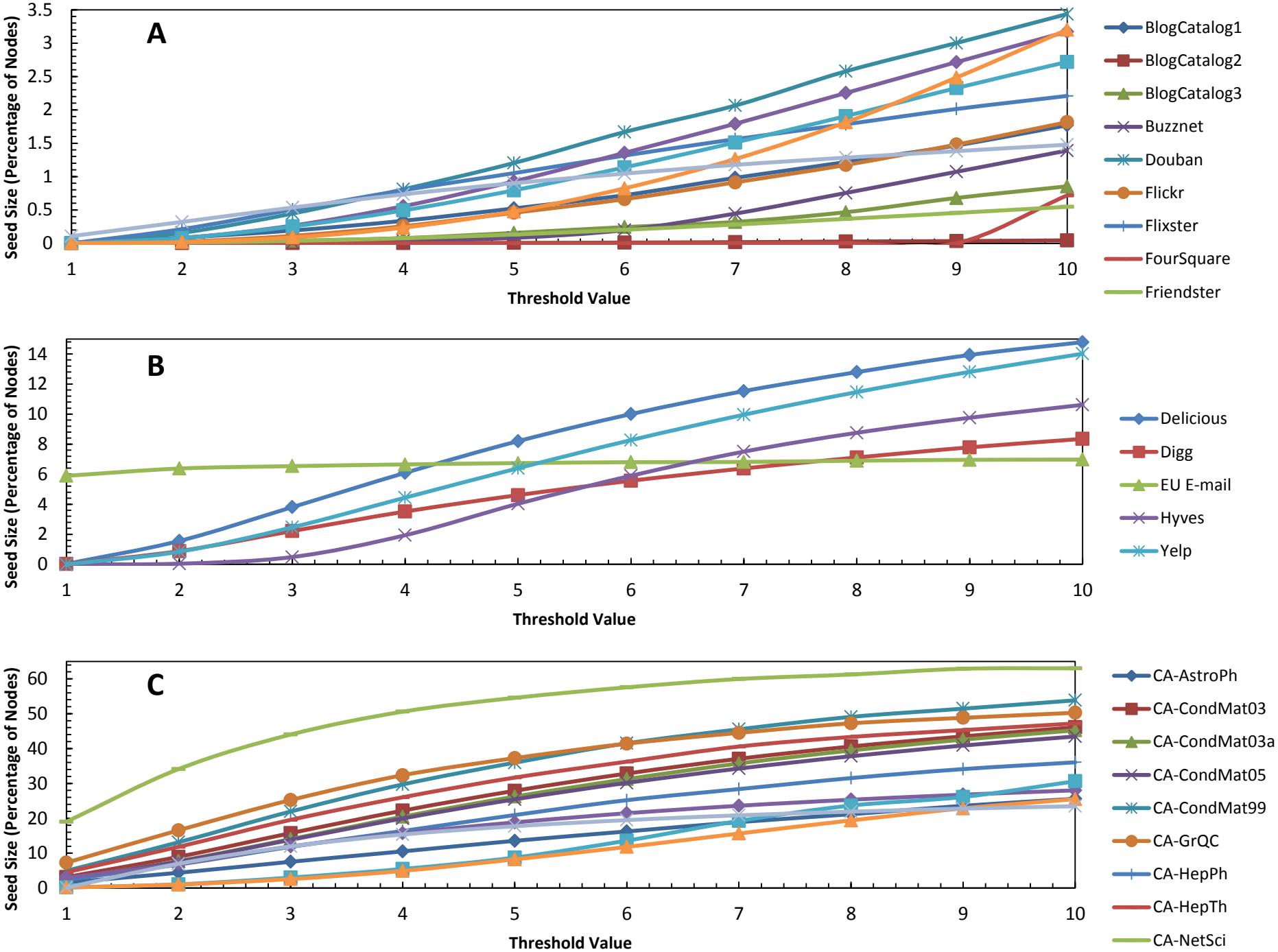
CATEGORY C

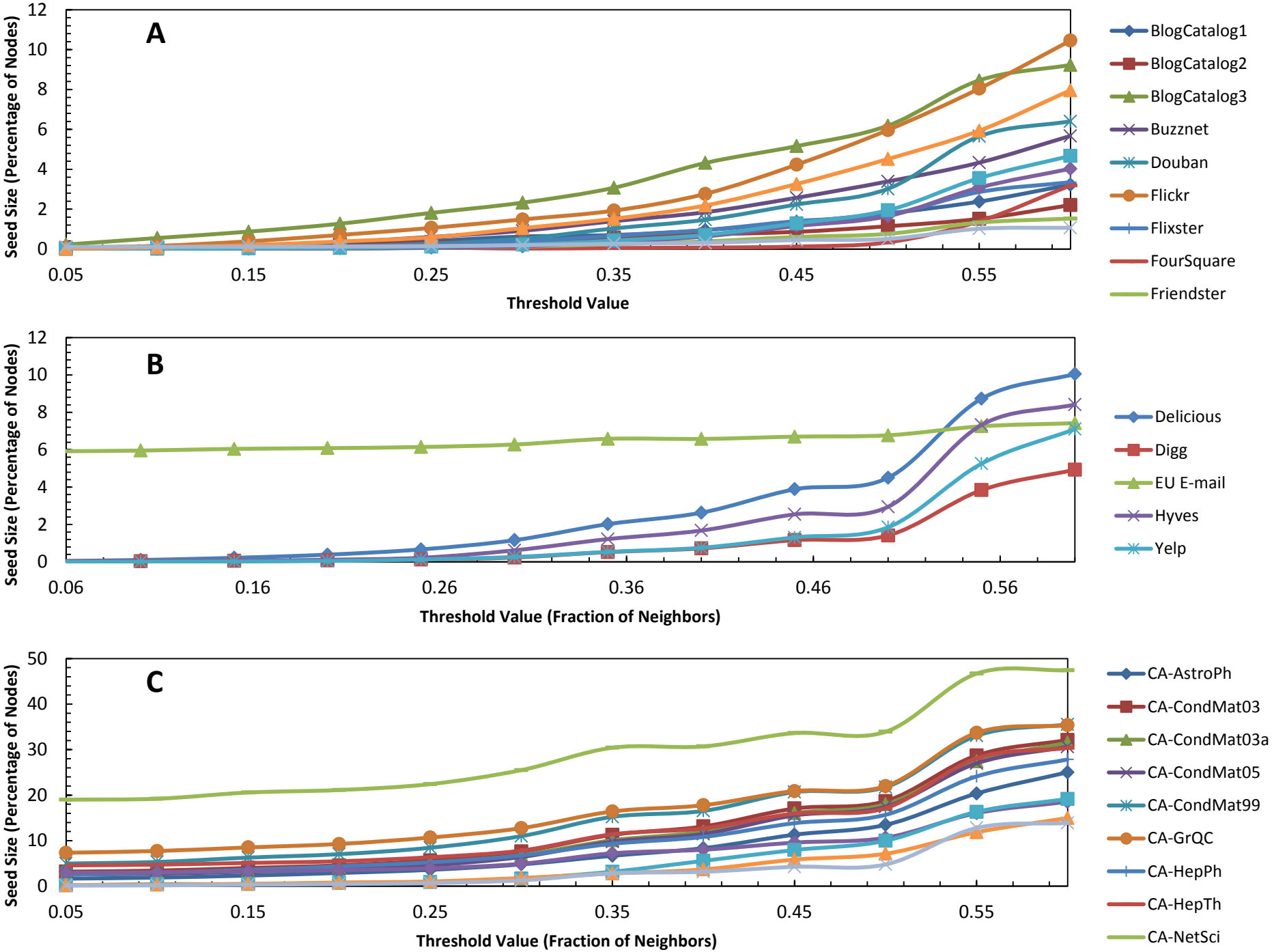
Name	# Nodes	# Edges	Avg. Degree	Source	Type
CA-AstroPh	18,772	396,100	10.55	SNAP	Collab
CA-CondMat03	30,460	240,058	3.94	UMICH	Collab
CA-CondMat03a	23,133	186,878	4.04	SNAP	Collab
CA-CondMat05	39,577	351,384	4.44	UMICH	Collab
CA-CondMat99	16,264	95,188	2.93	UMICH	Collab
CA-GrQc	5,242	28,968	2.76	SNAP	Collab
CA-HepPh	12,008	236,978	9.87	SNAP	Collab
CA-HepTh	9,877	51,946	2.63	SNAP	Collab
CA-NetSci	1,463	5,486	1.87	UMICH	Collab
Enron E-Mail	36,692	367,662	5.01	SNAP	E-Mail
URV E-Mail	1,133	10,902	4.81	URV	E-Mail
YouTube1	13,723	153,530	5.59	ASU	SocMedia
YouTube2	1,138,499	5,980,886	2.63	ASU	SocMedia

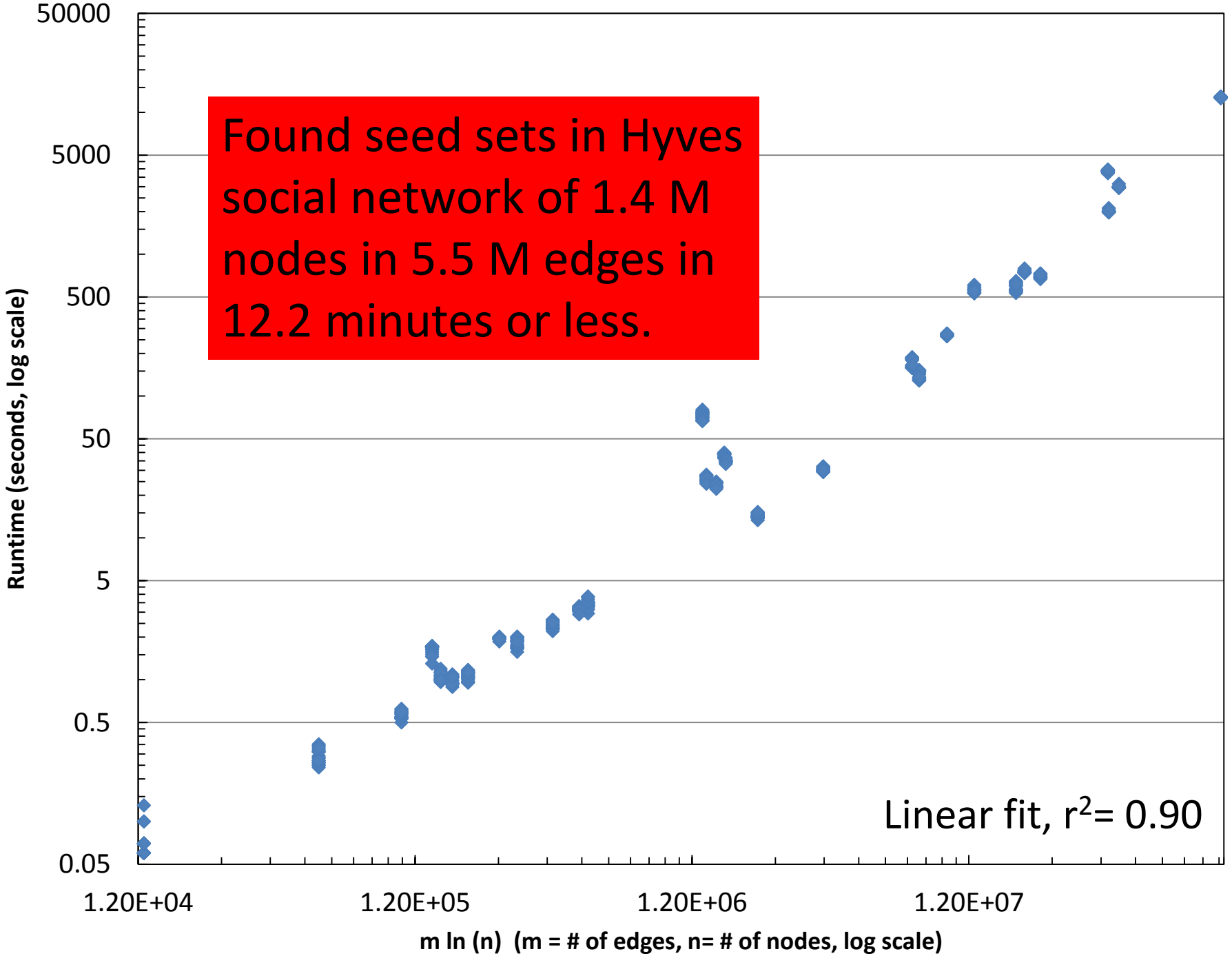
Main Results

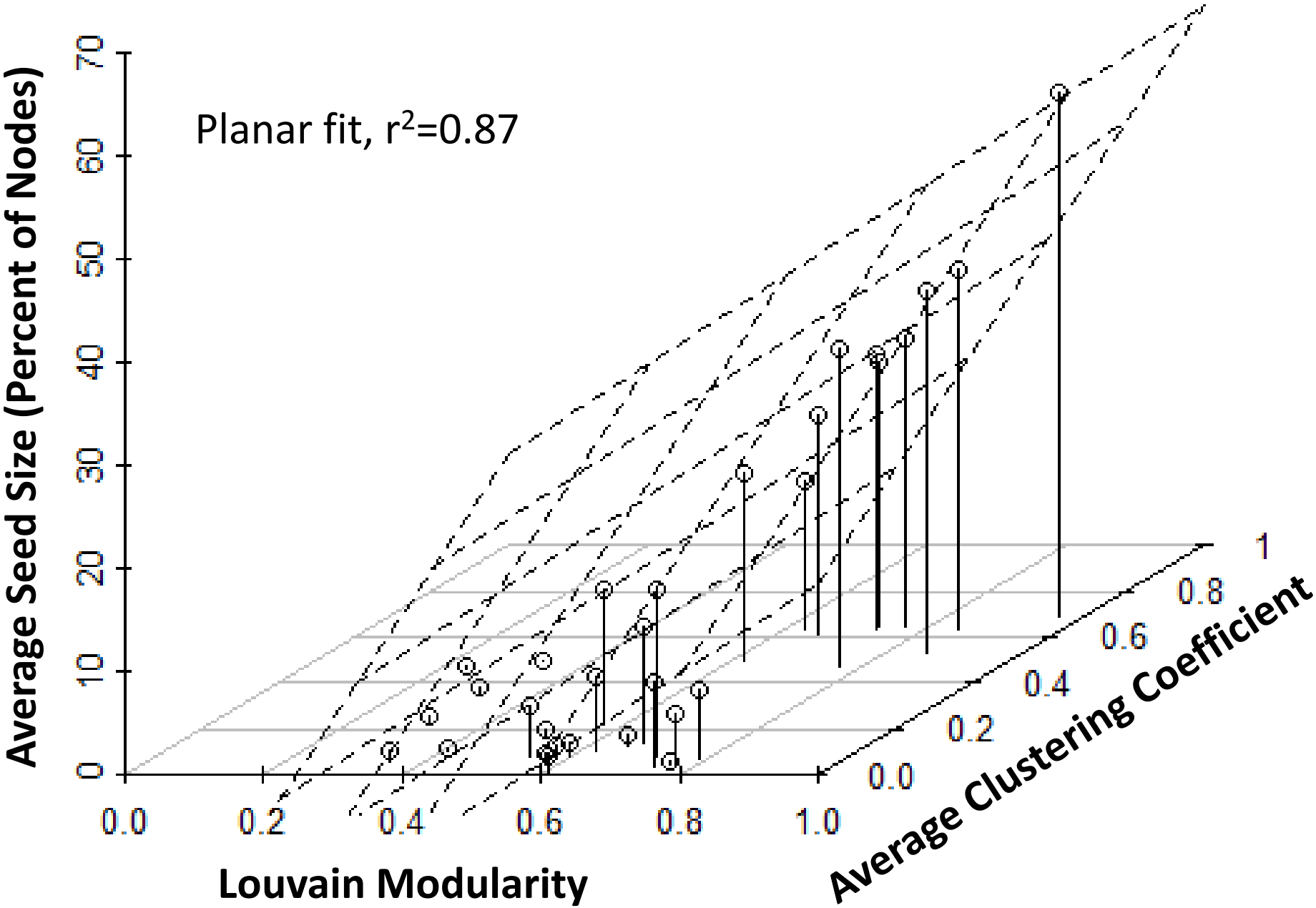
- Many social networks had very mall seed sets – many under 1% of the population, even for majority thresholds.
- The algorithm runs very quickly, able to process millions of nodes and edges in minutes.
- TIP_DECOMP tended to find small seed sets in networks with either a low average clustering coefficient and/or low Louvain modularity.

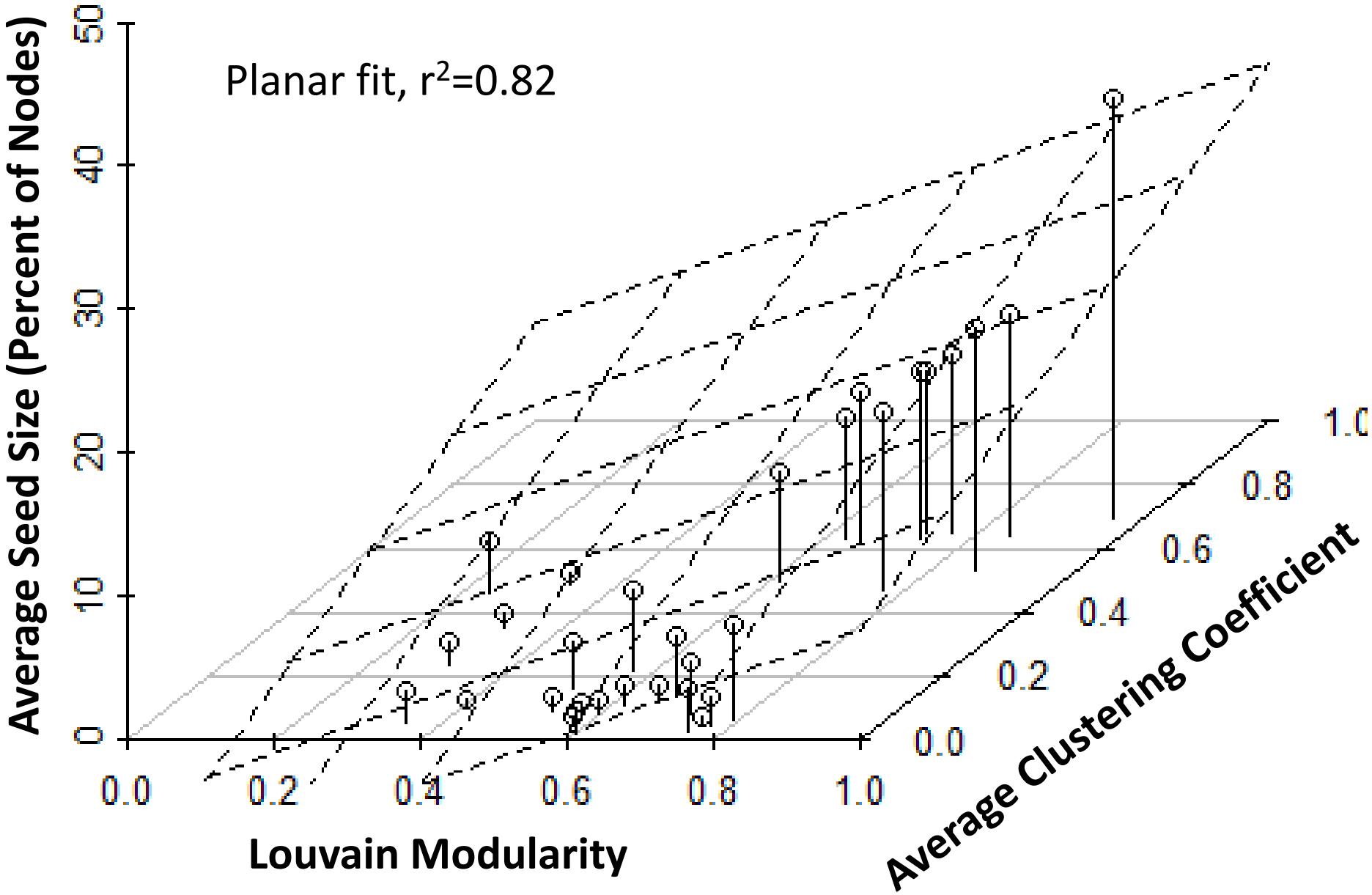
Network Science Center
West Point


A
Seed Size (Percentage of Nodes)
Threshold Value
BlogCatalog1
BlogCatalog2
BlogCatalog3
Buzznet
Douban
Flickr
Flixster
FourSquare
Friendster
B
Seed Size (Percentage of Nodes)
Threshold Value
Delicious
Digg
EU E-mail
Hyves
Yelp
C
Seed Size (Percentage of Nodes)
Threshold Value
CA-AstroPh
CA-CondMat03
CA-CondMat03a
CA-CondMat05
CA-CondMat99
CA-GrQC
CA-HepPh
CA-HepTh
CA-NetSci


A
Seed Size (Percentage of Nodes)
Threshold Value
0.05
0.15
0.25
0.35
0.45
0.55
BlogCatalog1
BlogCatalog2
BlogCatalog3
Buzznet
Douban
Flickr
Flixster
FourSquare
Friendster
B
Seed Size (Percentage of Nodes)
Threshold Value (Fraction of Neighbors)
0.06
0.16
0.26
0.36
0.46
0.56
Delicious
Digg
EU E-mail
Hyves
Yelp
C
Seed Size (Percentage of Nodes)
Threshold Value (Fraction of Neighbors)
0.05
0.15
0.25
0.35
0.45
0.55
CA-AstroPh
CA-CondMat03
CA-CondMat03a
CA-CondMat05
CA-CondMat99
CA-GrQC
CA-HepPh
CA-HepTh
CA-NetSci


Found seed sets in Hyves social network of 1.4 M nodes in 5.5 M edges in 12.2 minutes or less.
Linear fit, r^2= 0.90
50000
5000
500
50
5
0.5
0.05
Runtime (seconds, log scale)
1.20E+04
1.20E+05
1.20E+06
1.20E+07
m ln (n) (m = # of edges, n= # of nodes, log scale)


Planar fit, r2=0.87
Average Seed Size (Percent of Nodes)
0
10
20
30
40
50
60
70
Louvain Modularity
0.0
0.2
0.4
0.6
0.8
1.0
Average Clustering Coefficient
0.0
0.2
0.4
0.6
0.8
1


Planar fit, r2=0.82
Average Seed Size (Percent of Nodes)
0
10
20
30
40
50
Louvain Modularity
0.0
0.2
0.4
0.6
0.8
1.0
Average Clustering Coefficient
0.0
0.2
0.4
0.6
0.8
1.0

Concluding Remarks

- TIP_DECOMP is a simple, but powerful algorithm for finding seed sets in a social network under the tipping model.
- Experimental evaluation shows that it can easily scale to networks with millions of nodes and edges.
- TIP_DECOMP tends to find smaller seed sets in networks with less community structure. Fortunately, many online social networks seem to exhibit this quality.

Network Science Center
West Point

Questions

Network Science Center
West Point