Algorithmic Network Science

6th Annual Network Science Workshop
West Point, NY
23 April 2012

Paulo Shakarian, Ph.D.
Major, U.S. Army
Dept. Electrical Engineering and Computer Science and Network Science Center
U.S. Military Academy, West Point NY

Network Science Center
West Point

Algorithmic Network Science

The intersection between network science and algorithmic design with the objective of designing algorithms to solve or approximate hard network science problems from a variety of disciplines.

Network Science Center
West Point

The Algorithmic Network Science Group at the West Point NSC (Faculty)

- MAJ Paulo Shakarian, Ph.D.
- EECS
  - LTC Christa Chewar, Ph.D.
  - LTC Paul Maxwell, Ph.D.
  - MAJ Devon Callahan
  - MAJ Jesse Fleming
  - CPT(P) Ryan Tate
- SE
  - MAJ Charles Levine
  - James Cook
- MATH
  - MAJ Anthony Johnson, Ph.D.
  - Dr. Brian MacDonald
  - MAJ Nick Howard
- LIFE
  - Dr. J. Kenneth Wickiser
- External
  - Dr. Gerardo I. Simari, Oxford University
  - Patrick Roos, UMD

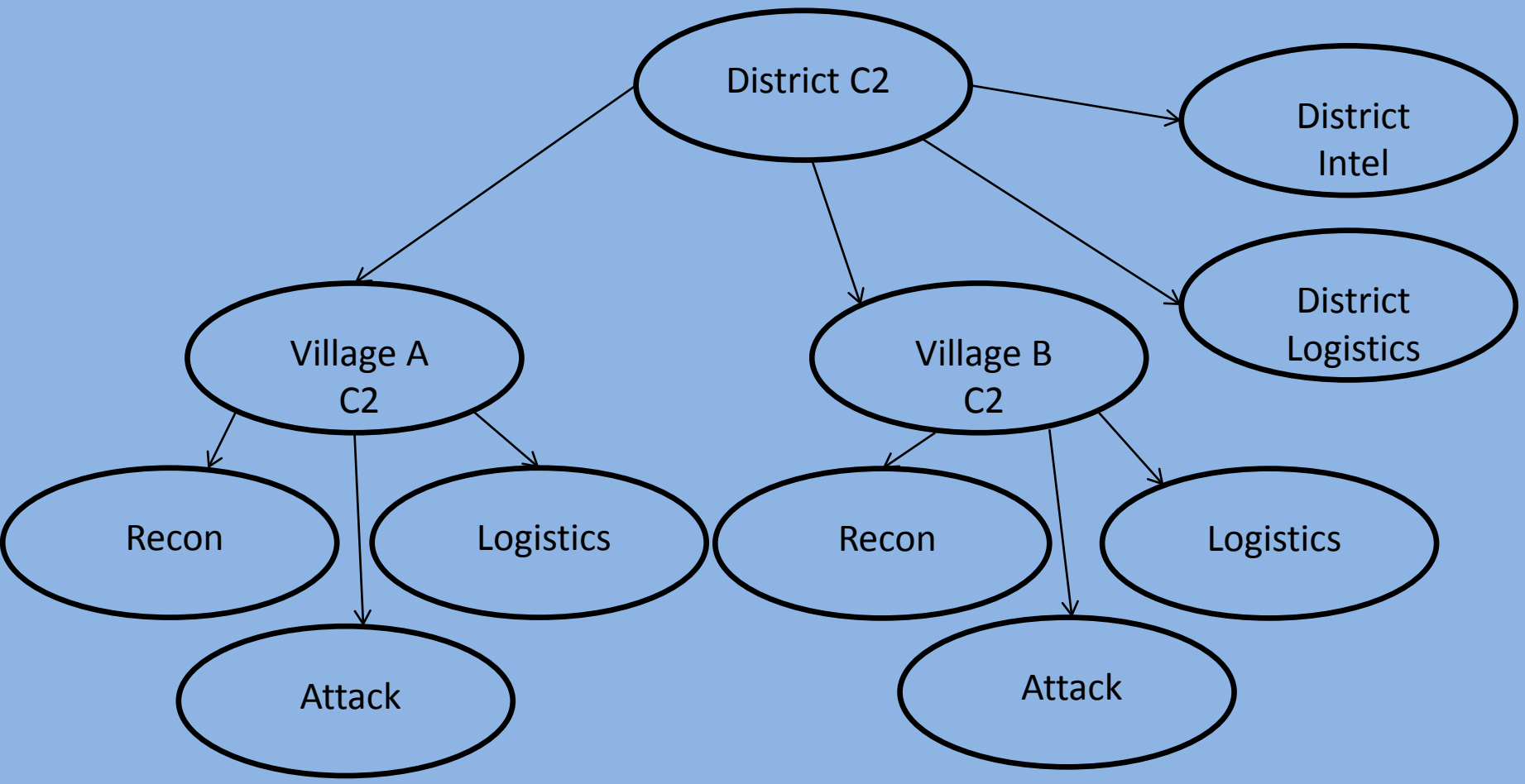
Network Science Center
West Point

The Algorithmic Network Science Group at the West Point NSC (Cadets)

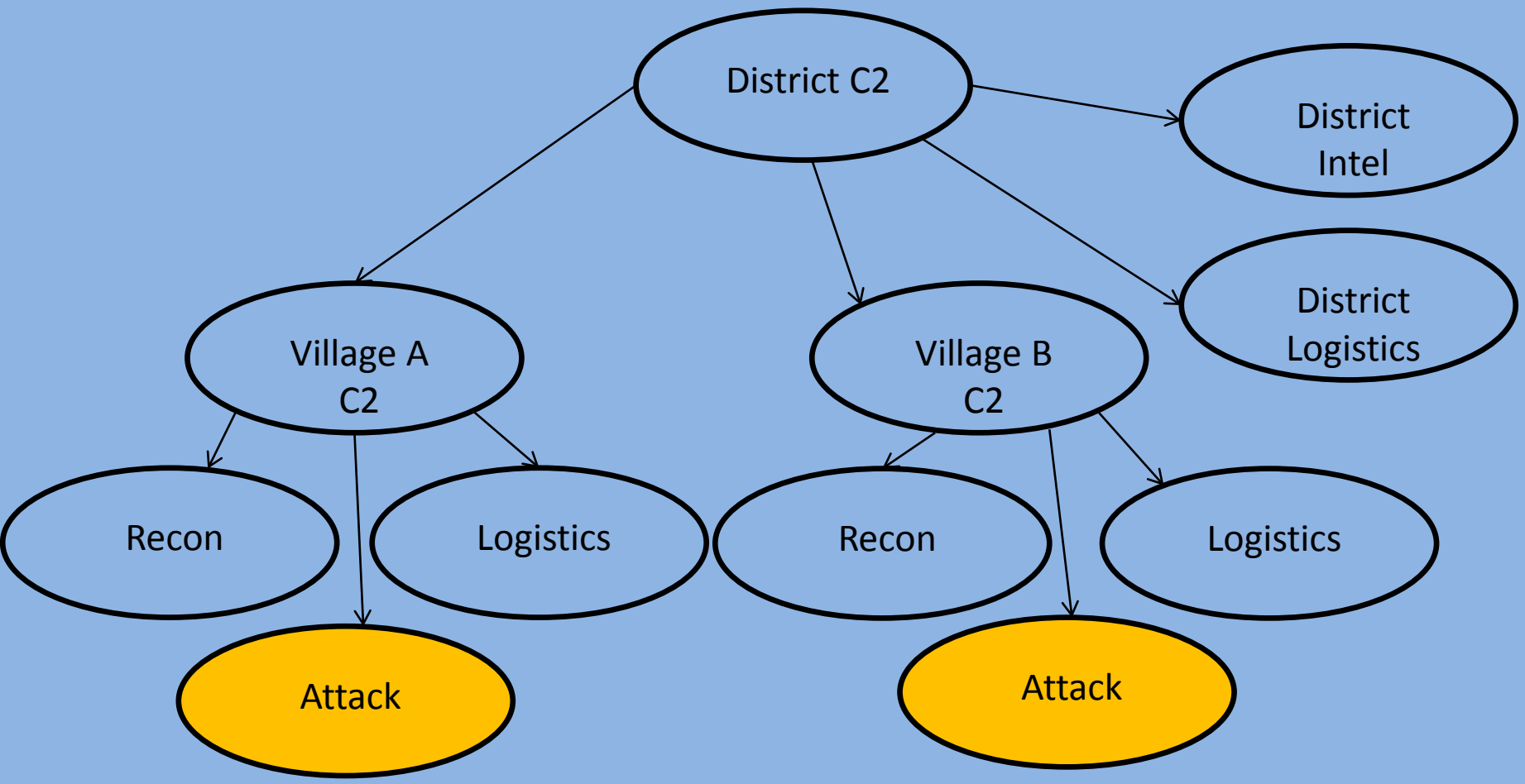
- EECS
  - CDT Sean Eyre
  - CDT Randolph Rotte
  - CDT Theodore Taggart
  - CDT Andrew Oswald
  - CDT Ellis Valdez
- MATH
  - CDT Jeffrey Nielsen
  - CDT Damon Paulo
- SE
  - CDT Calvin Santos
  - CDT Patrick Lockard
  - CDT Zachary Bloom
  - CDT Jason Newell
- ECON
  - CDT Chad Vessel
- Physics
  - CDT Geoffrey Moores

Network Science Center
West Point

Consider this model of an insurgent network…

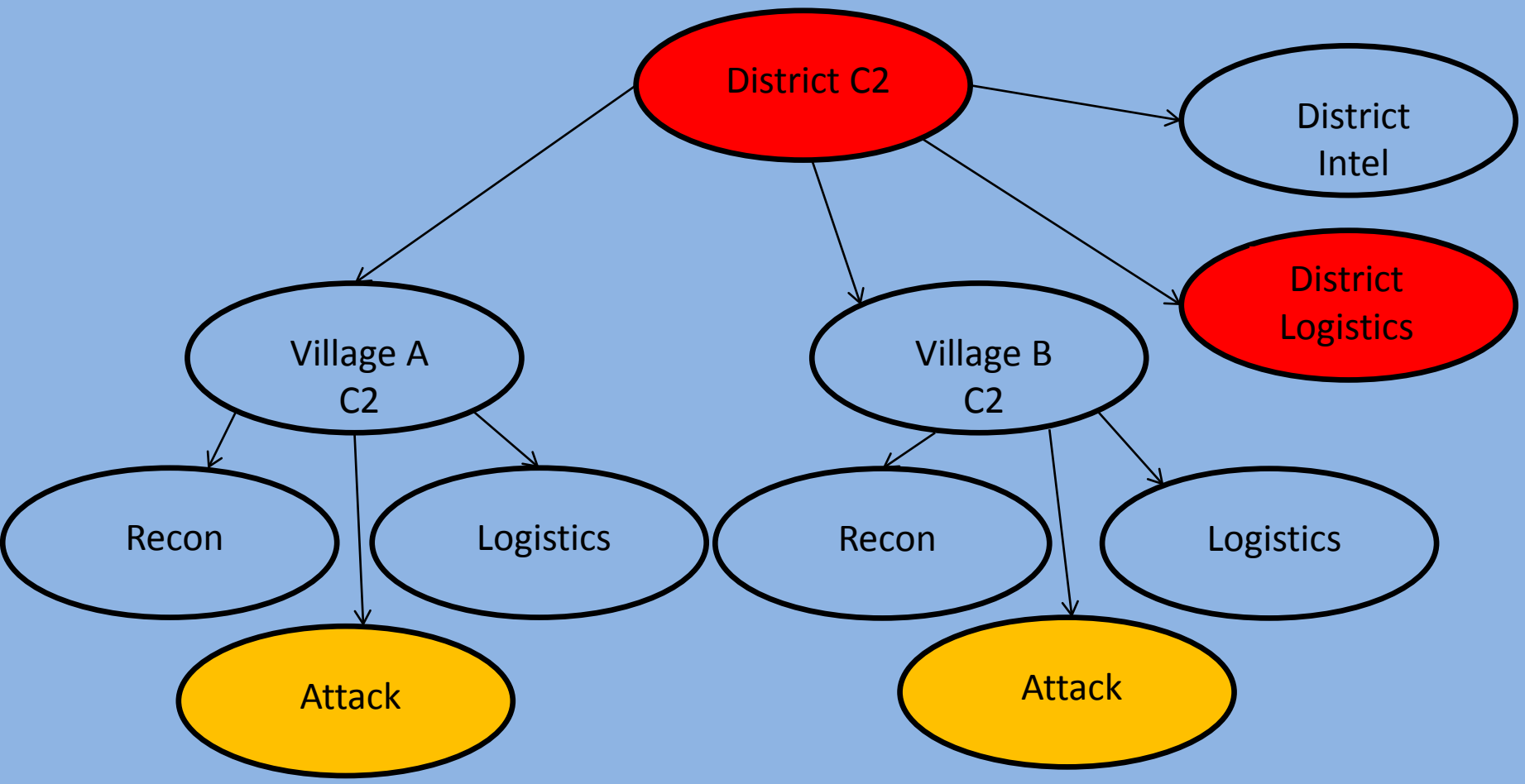


Consider this model of an insurgent network…



If we know where the attacks are occurring,

Consider this model of an insurgent network…

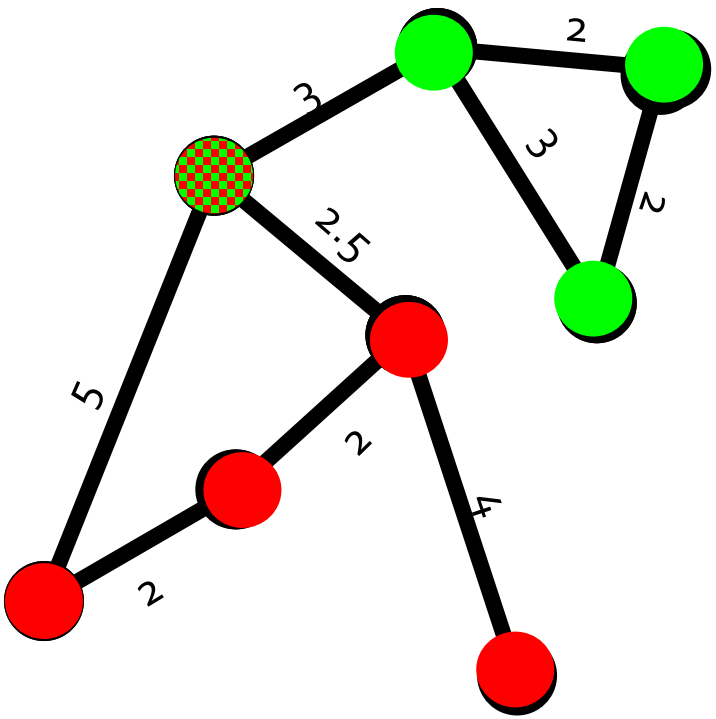


If we know where the attacks are occurring,
Can we locate insurgent command and control and depot-level logistics sites (a.k.a. high value targets)?

Village Relationship Network (VRN)

We use a "tribal relationship network" to model the relationship among cells.

- Network with 1,543 vetices representing villages in Helmand, Kandahar, and Zabul
- 1 Million edges among the vertices, present if there exists a tribal relationship between the two villages and weighted by the road distance between the two

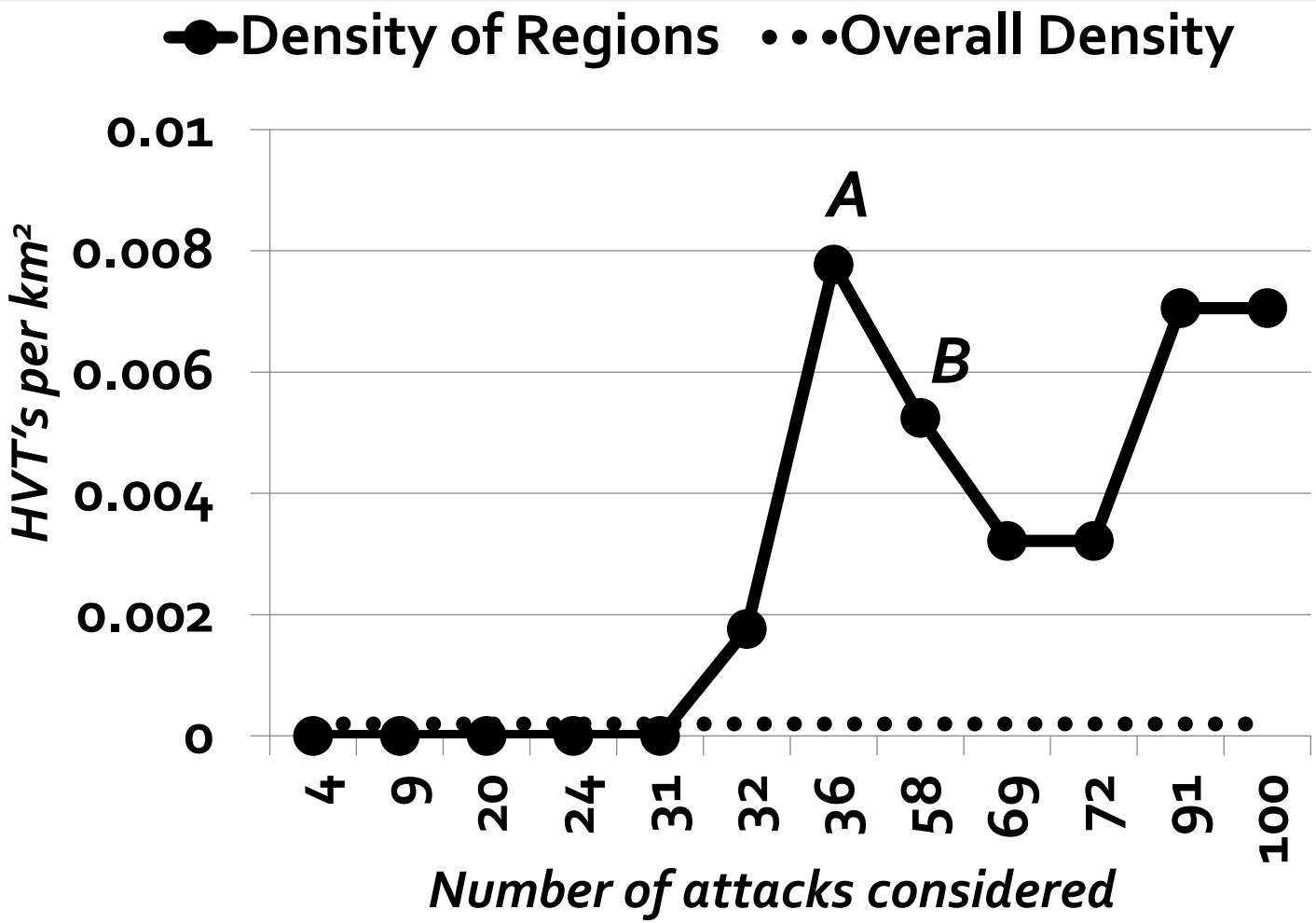


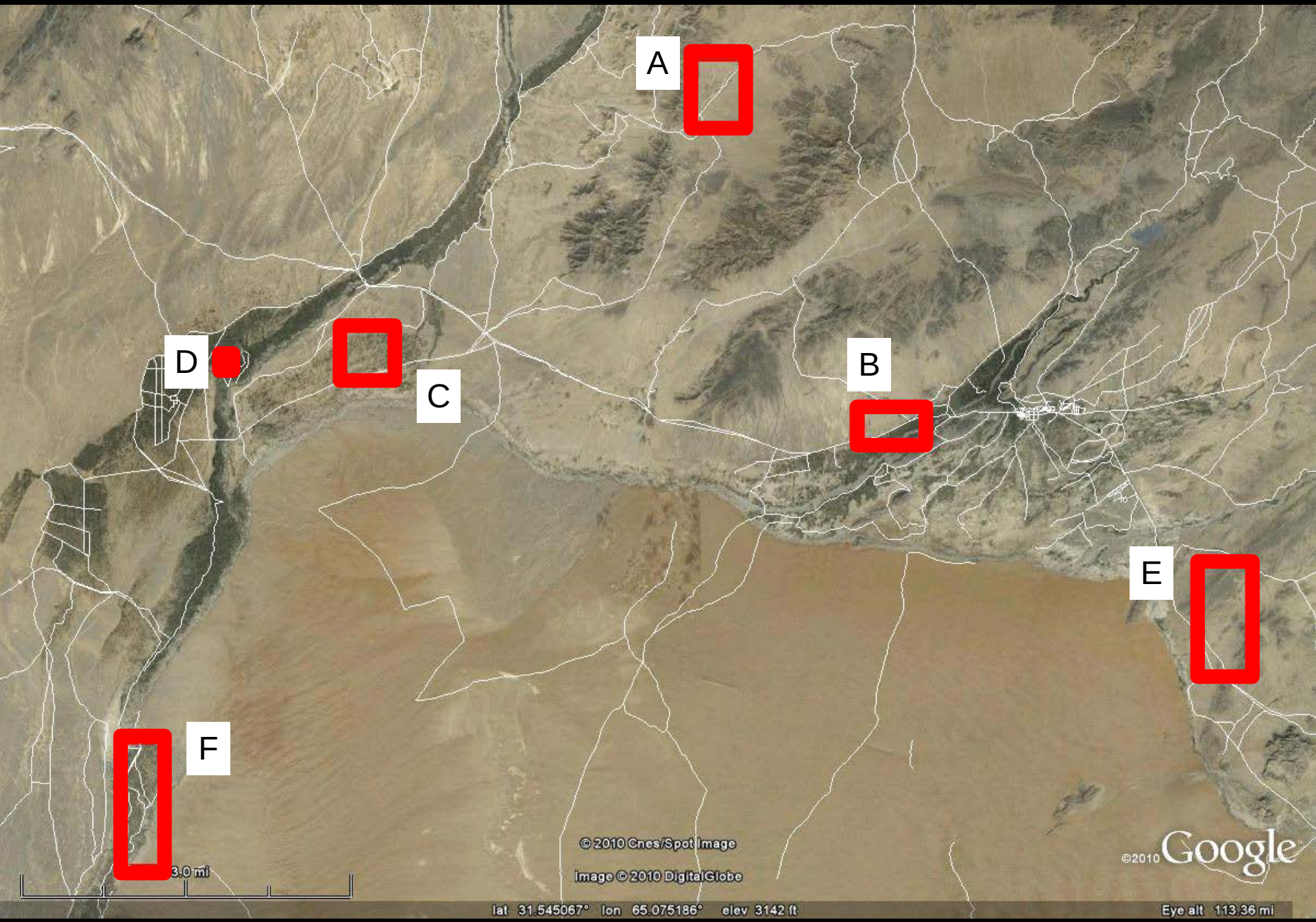
Results

Leveraging ***geospatial abduction***, the SCARE-S2 software found minimally enclosing rectangles around villages of ~100 km^2 that were 33x more likely to contain a high value targets.

Classified tests were conducted at ARL ALC.

HVT Density of Rectangles Returned by SCARE-S2




A
B
C
D
E
F
© 2010 Cnes/Spot Image
Image © 2010 DigitalGlobe
©2010 Google
3.0 mi
lat 31.545067° lon 65.075186° elev 3142 ft
Eye alt 113.36 mi

Extending the VRN

Currently, the VRN uses edges that are weighted by the shortest path length on a road network.

Can we consider travel time?

Can we consider road conditions?

Can we consider vehicle type?

Yes, using *Common Ground*, we can create new VRN's where the edges can be weighted using the above ideas.

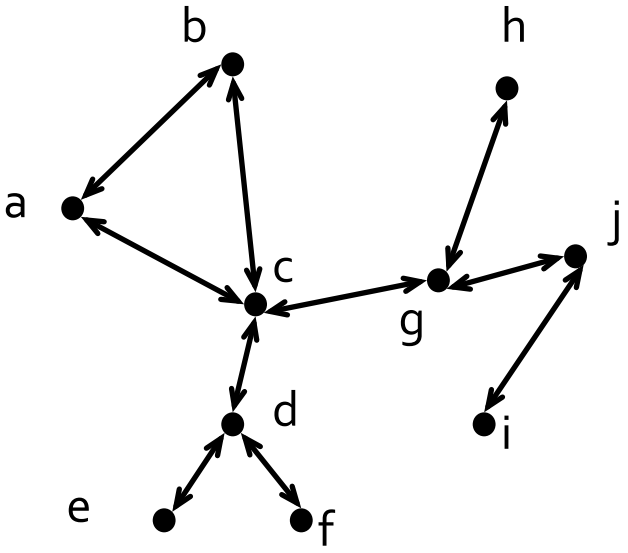
Classified tests to occur this summer as part of our effort to support of the 588th Engineer Detachment.

Network Science Center
West Point

How else can a VRN be Leveraged?

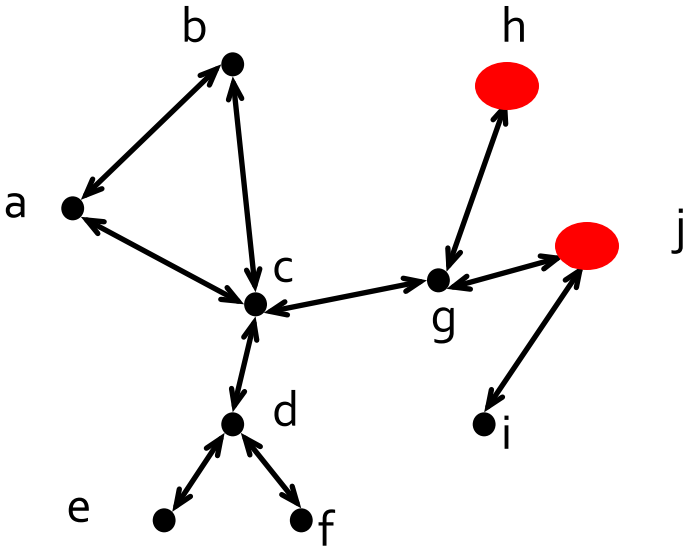
Suppose we use "tipping" to model the spread of anti-insurgent sentiment.

Suppose a village node "tips" to anti-insurgent behavior when half of the adjacent villages have that behavior.



How else can a VRN be Leveraged?

Initially villages **h** and j are anti-insurgent.



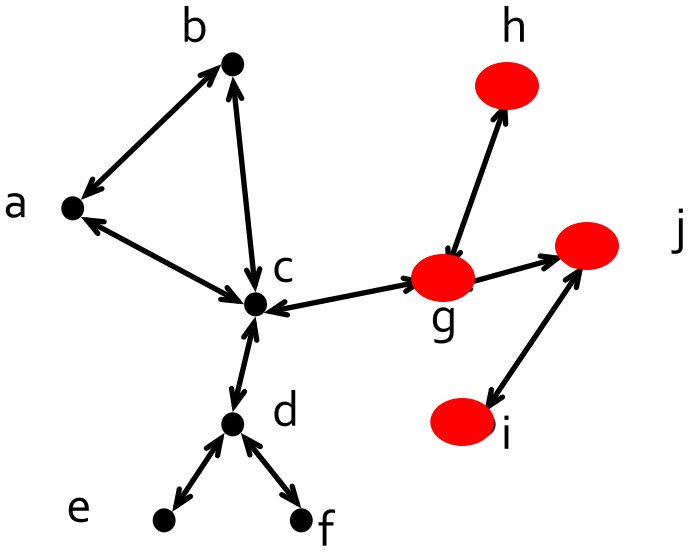
Network Science Center
West Point

How else can a VRN be Leveraged?

Initially villages **h** and **j** are anti-insurgent.

This leads to **g** and **i** also becoming anti-insurgent.

But no others.



Extending the VRN

In 2010, Special Forces trained locals of the village of Gizab who revolted against the Taliban.

If we assume a tipping model on the VPN, perhaps we can identify a set of nodes that causes all the villages to take anti-Taliban action.

Hence, we ask the following question:
Can we find a minimal set of nodes under the tipping model that ensures that the entire population adopts the behavior in question?

This the *minimum-seed* problem.

Network Science Center
West Point

The Minimum Seed Problem

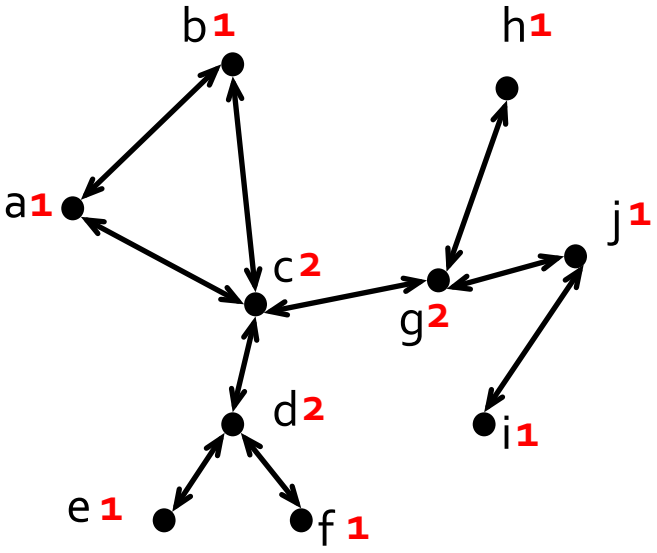
- Minimum-seed is NP-complete.
- Our new pruning heuristic guarantees that the result will lead to adoption of the behavior by the entire population, but does not guarantee minimal size.
- However, experimentally, we find that the set returned is often several orders of magnitude smaller than the population.

Network Science Center West Point

The Pruning Heuristic for the Tipping Model

Let us consider the following network.

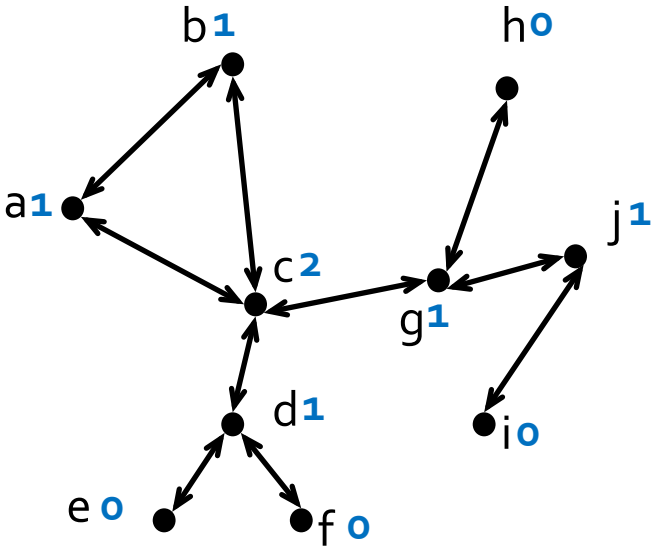
First, our algorithm identifies the threshold for each node to be tipped. If we assume at least 50% of each neighbor is required, then this number is shown in red.



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

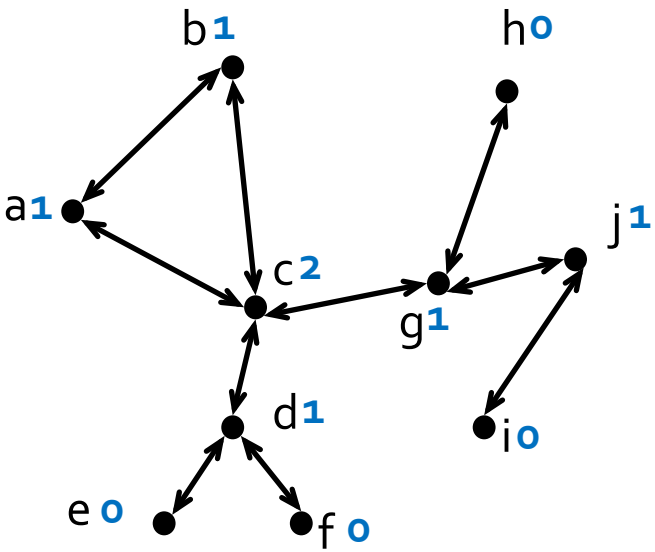


The Pruning Heuristic for the Tipping Model

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.

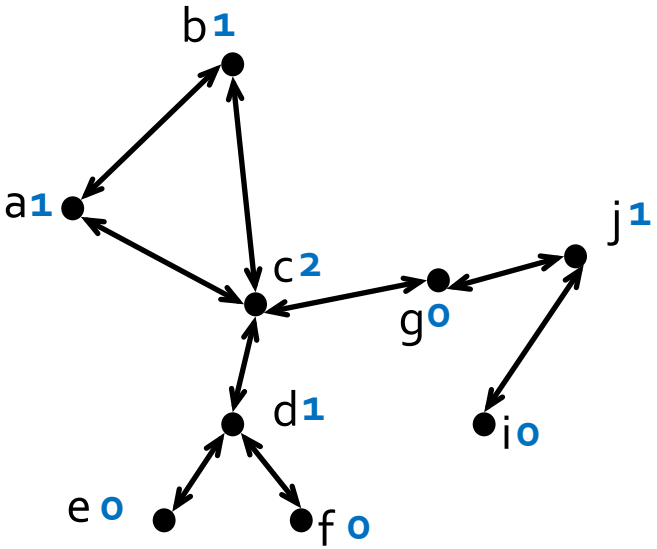


The Pruning Heuristic for the Tipping Model

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.

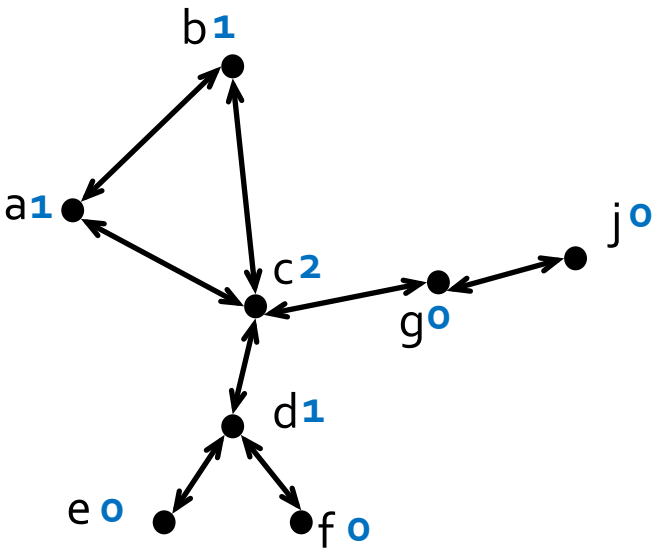


The Pruning Heuristic for the Tipping Model

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.

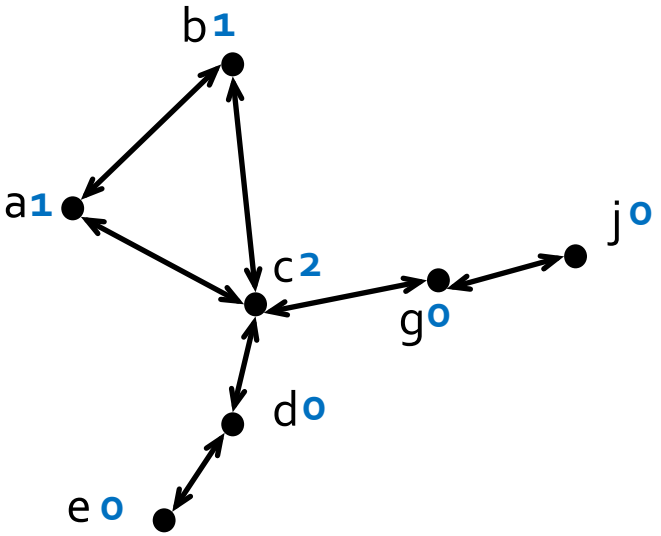


The Pruning Heuristic for the Tipping Model

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.

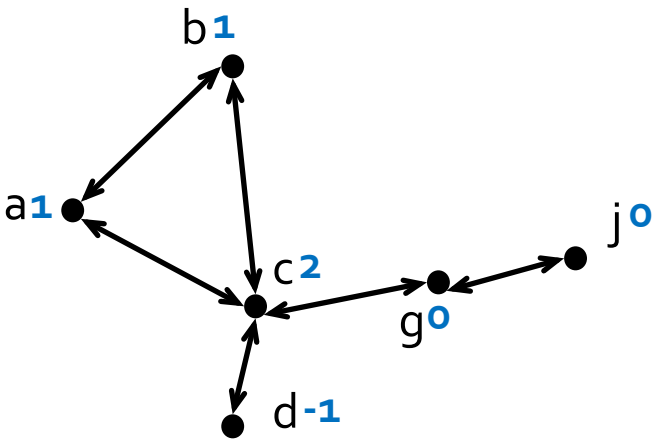


The Pruning Heuristic for the Tipping Model

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.

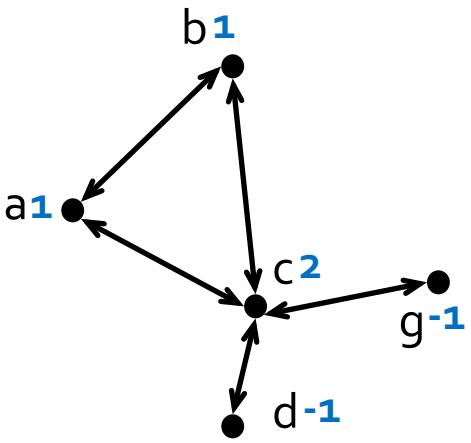


The Pruning Heuristic for the Tipping Model

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.

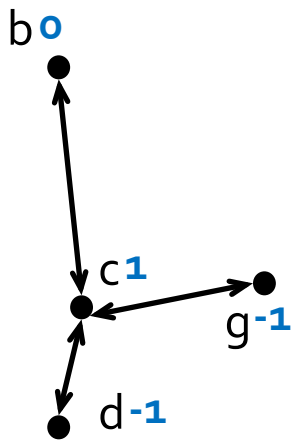


The Pruning Heuristic for the Tipping Model

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.

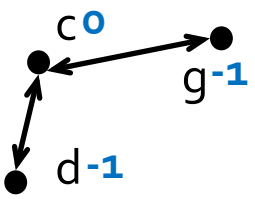


The Pruning Heuristic for the Tipping Model

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.

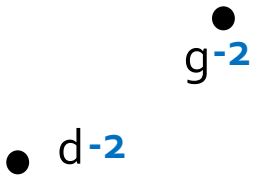


The Pruning Heuristic for the Tipping Model

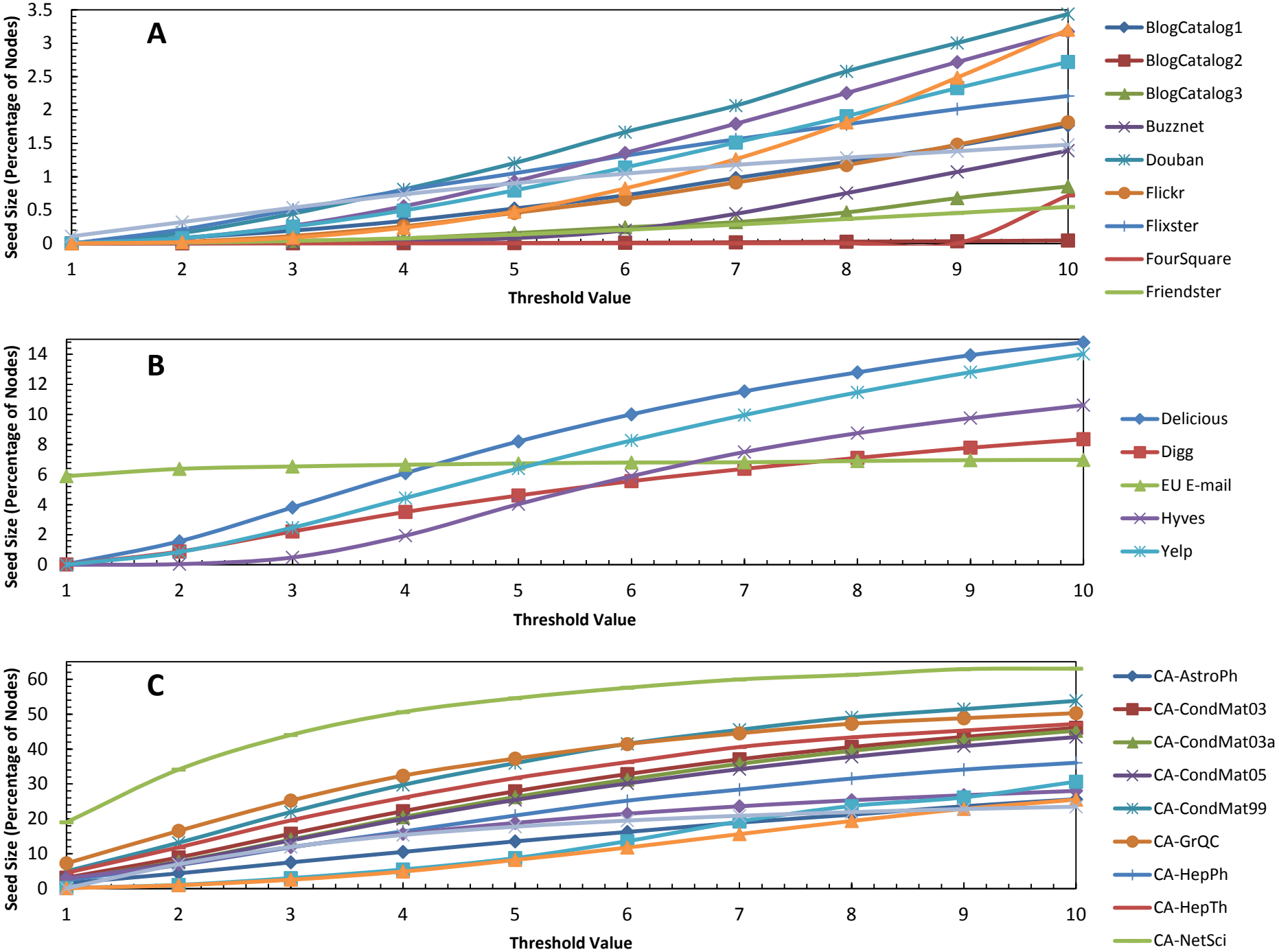
The algorithm then proceeds as follows:

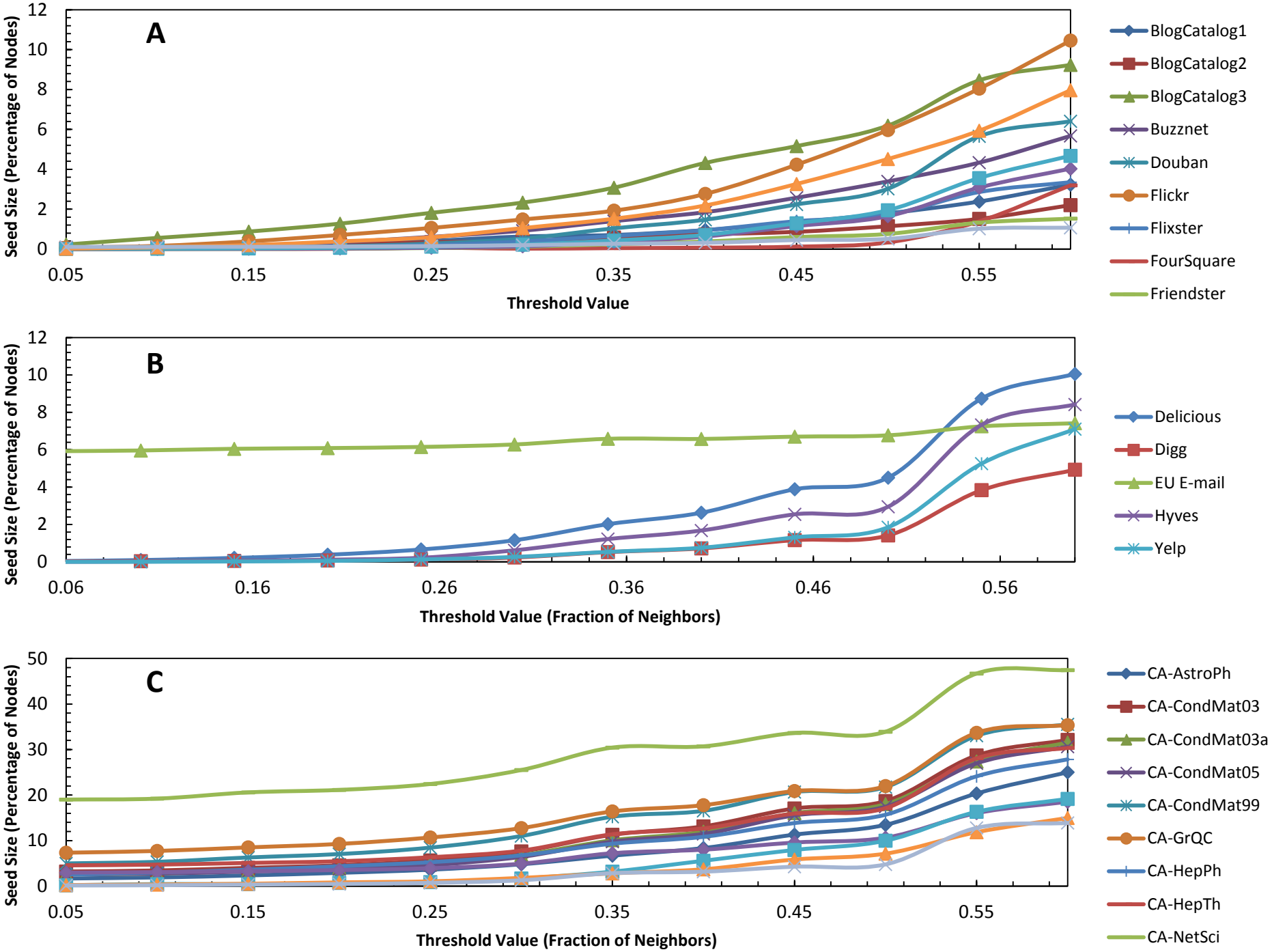
At each iteration, remove the node whose distance is the smallest but non-negative (and the adjacent edges)

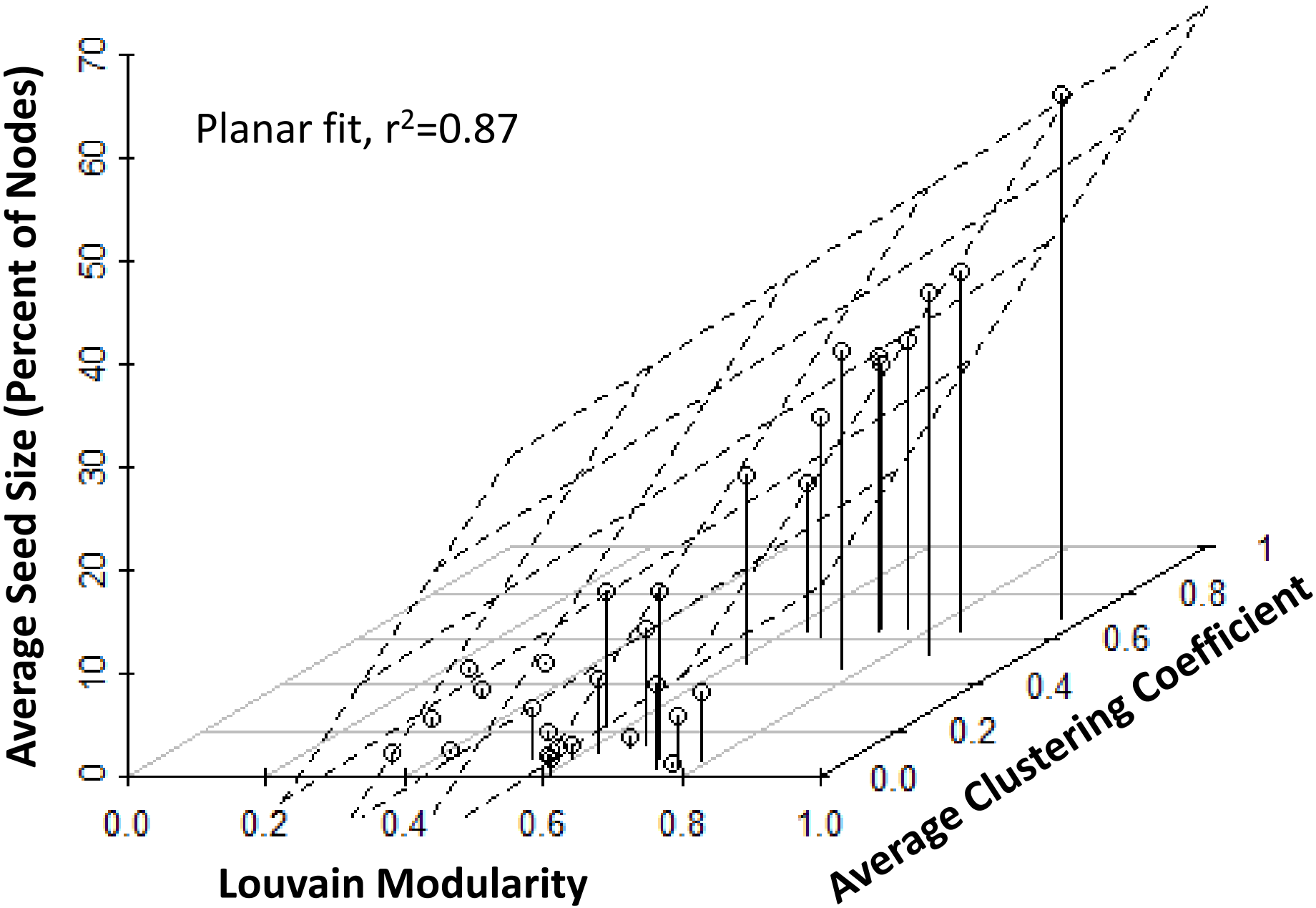
Then update the distances of its neighbors.

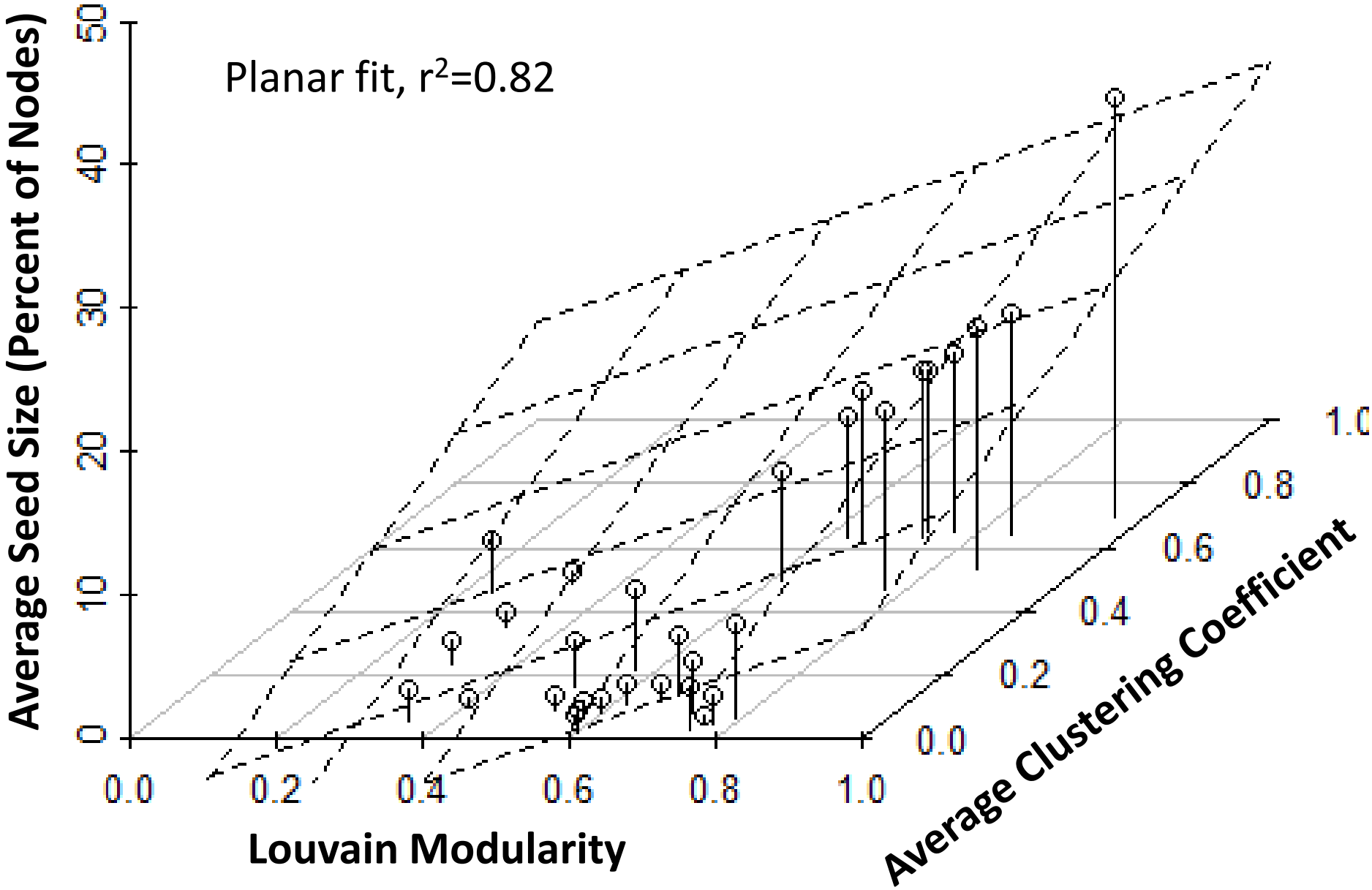


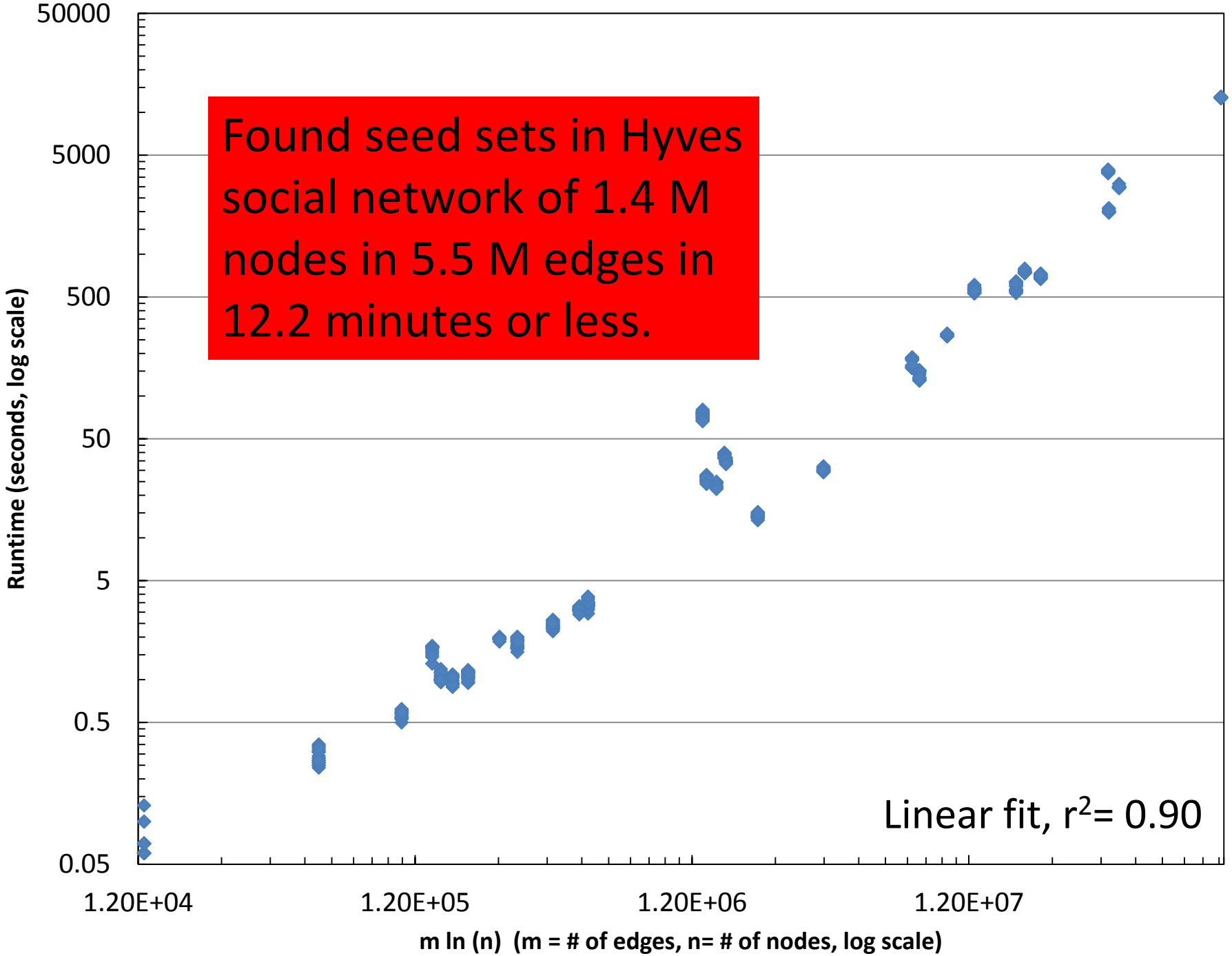
Solution = { d, g }


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


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


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)

Other Applications of Pruning in Networks

What other problems can pruning help us solve?

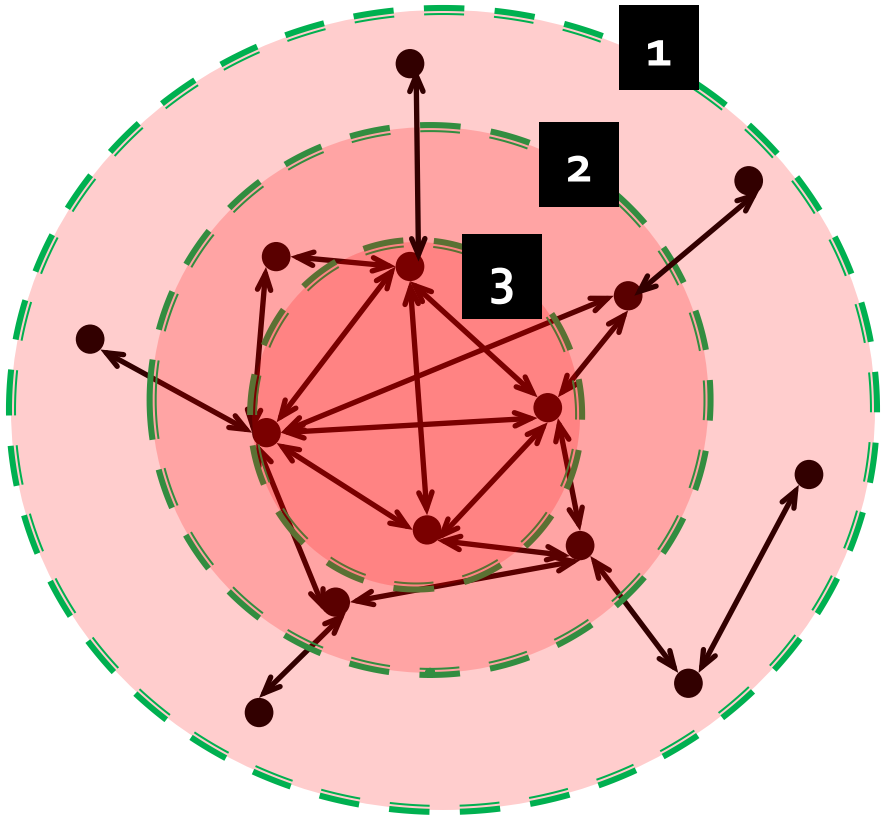
- Consider the protein networks of an organism and a pathogen attacking that organism.
- *Can we characterize the nodes that are most attacked in the organism?*
- We use shell-decomposition to study this problem.

Shell Decomposition

The k-core of a network is a subgraph where each node is connected to the rest of the subgraph by at least k edges.

The k-shell is the set of nodes in a k-core but not the next higher k-core.

The k-shells of a network can be identified with a simple pruning process.



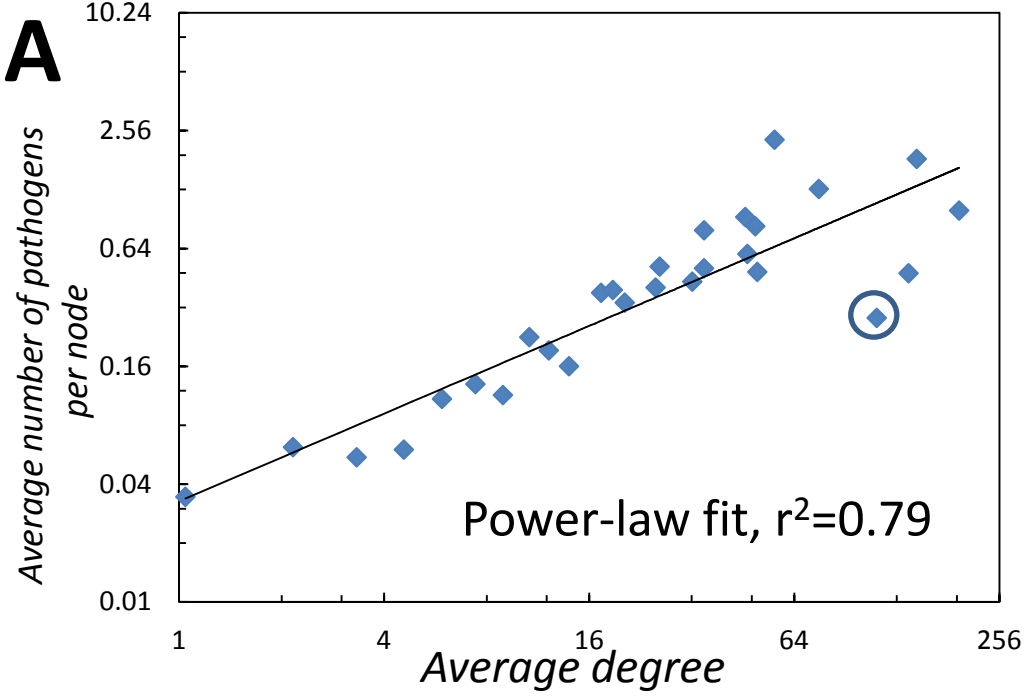
Characterizing Proteins Targeted by Pathogens Using Centrality

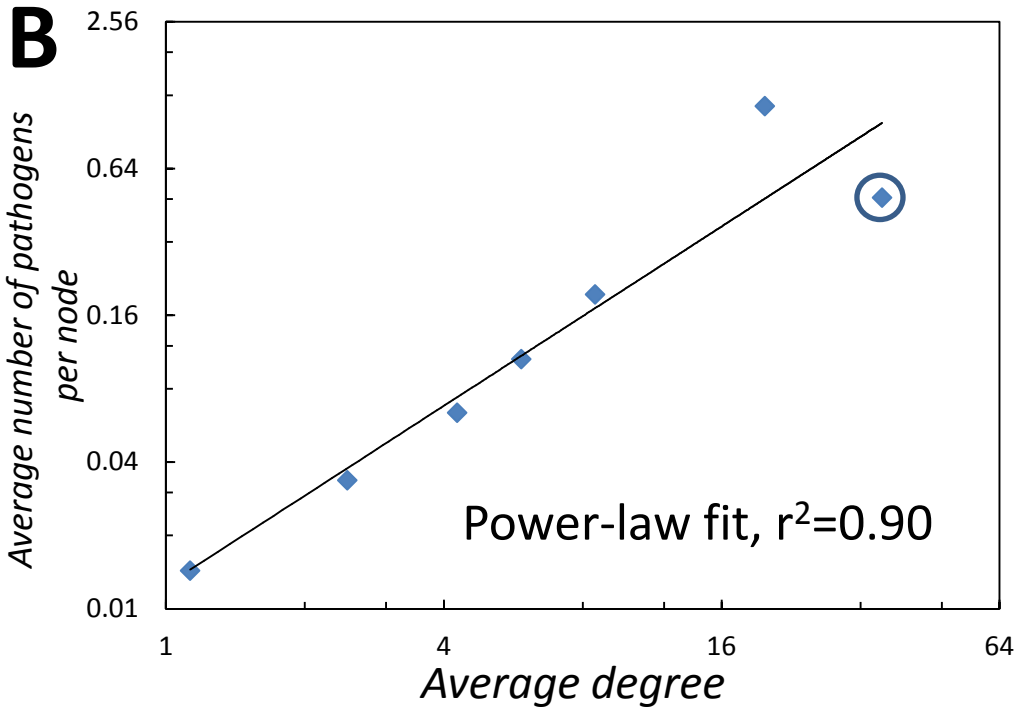
The centrality measures from the literature correlate little with the number of pathogens targeting a node, in a human protein network, we obtained the following:

- Degree: r^2=0.06 , MIC= 0.08
- Betweenness: r^2=0.04 , MIC= 0.11
- Shell number: r^2=0.03 , MIC= 0.07

Combining Decomposition and Centrality

We looked at the different shells of the protein networks and compared the *average degree* of the shell with the *average number of pathogen effectors* for that shell and found strong correlations for both humans and the *Arabidopsis thaliana* reference plant.

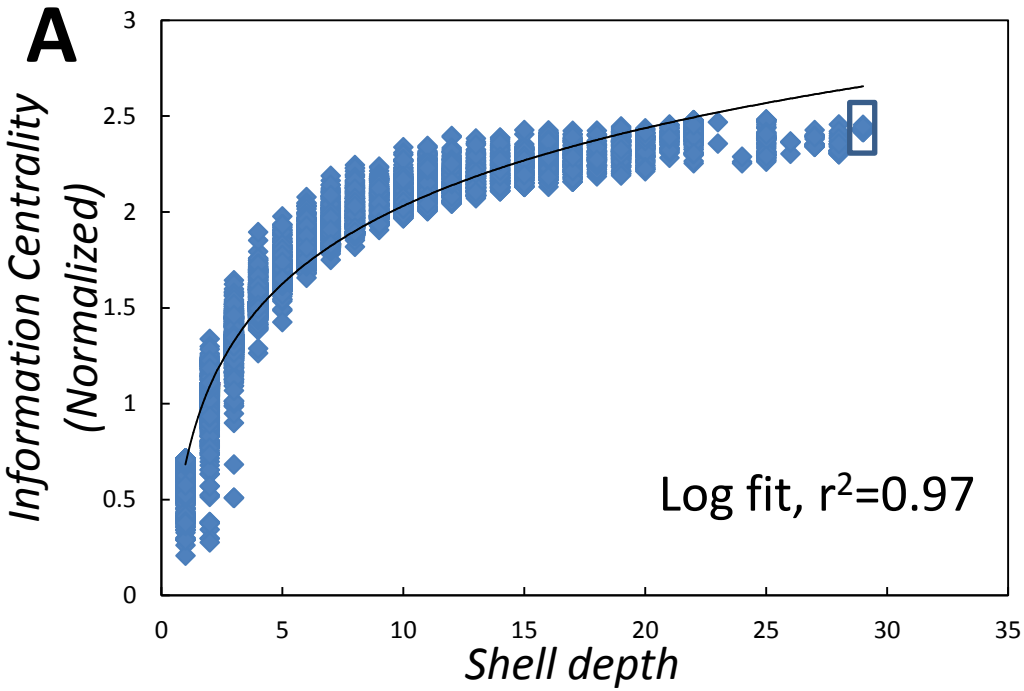


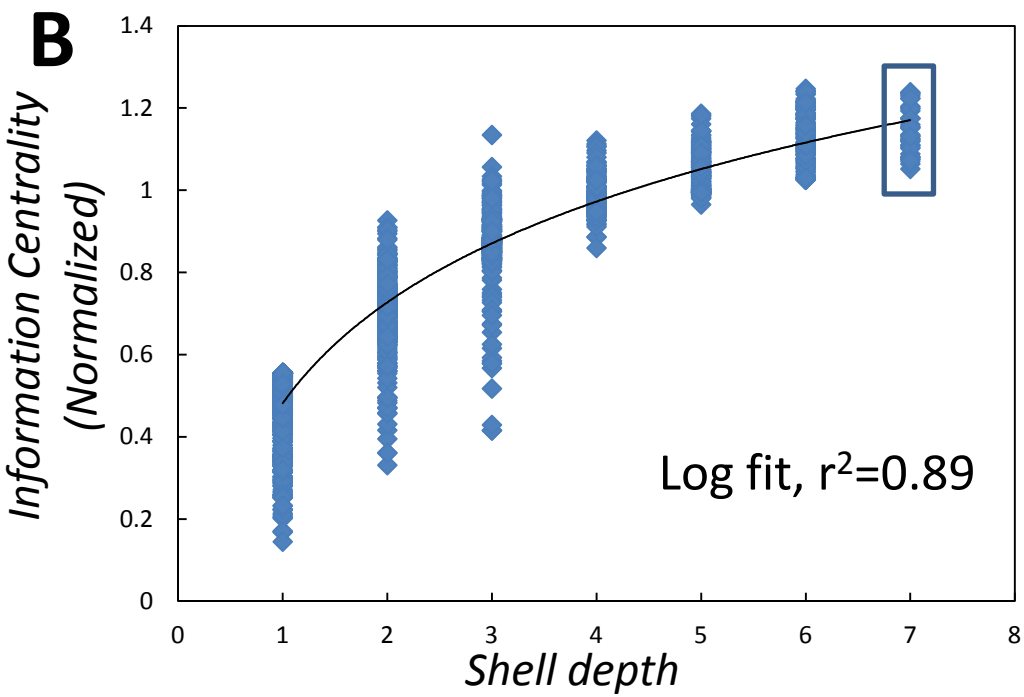


We find a strong power-law correlation between average shell degree and average number of pathogen attacks.

However, core nodes (circled) are significantly under-targeted.

- Human protein network shown in panel A.
- Arabidopsis protein network shown in panel B.





We also found a log-relationship between information centrality and shell-depth (core nodes are boxed).

This may indicate that core nodes are key spreaders of intra-cellular communication.

Pathogens may attack these nodes less to ensure the survival of the host.

- Human protein network shown in panel A.
- Arabidopsis protein network shown in panel B.

Summary

Our new team dedicated to Algorithmic Network Science is tackling a variety of hard problems. Today we discussed the following:

- Finding high-value targets in Afghanistan
- Finding seed sets in social networks
- Identifying proteins likely targeted by pathogens

Network Science Center
West Point

References

- SCARE-S2:
  - Shakarian, P. and Subrahmanian, V.S., *Geospatial Abduction Principles and Practice*, Springer, NY 2011.
- Pruning heuristic for the Minimum Seed Problem:
  - Shakarian, P. and Paulo, D., *Large Social Networks can be Targeted for Viral Marketing with Small Seed Sets*, submitted 2012.
- Pathogens and protein networks:
  - Shakarian, P. and Wickiser, J.K., *Similar Pathogen Targets in Arabidopsis thaliana and Homo sapiens Protein Networks*, submitted 2012.

Network Science Center West Point

Questions

Network Science Center
West Point