KENT STATE UNIVERSITY

Information Diffusion on Social Networks

Zhuo Chen
Lanxue Dang
Xinyue Ye
Jay Lee

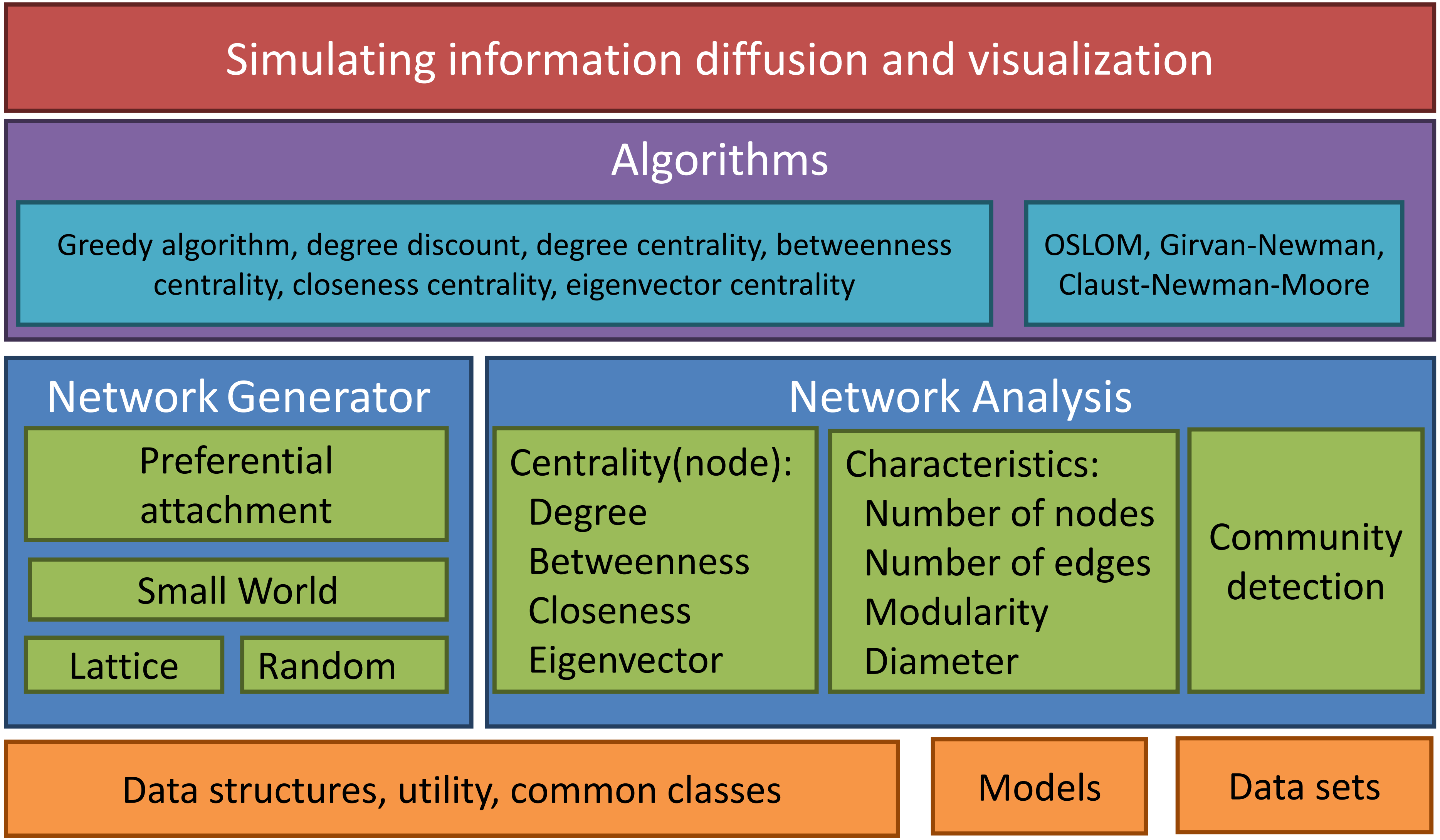
Outline

1. **Introduction**
2. Social Network Tools and Algorithms
3. Simulation and Analysis of Information diffusion
4. Concluding Remarks
5. Next Tasks

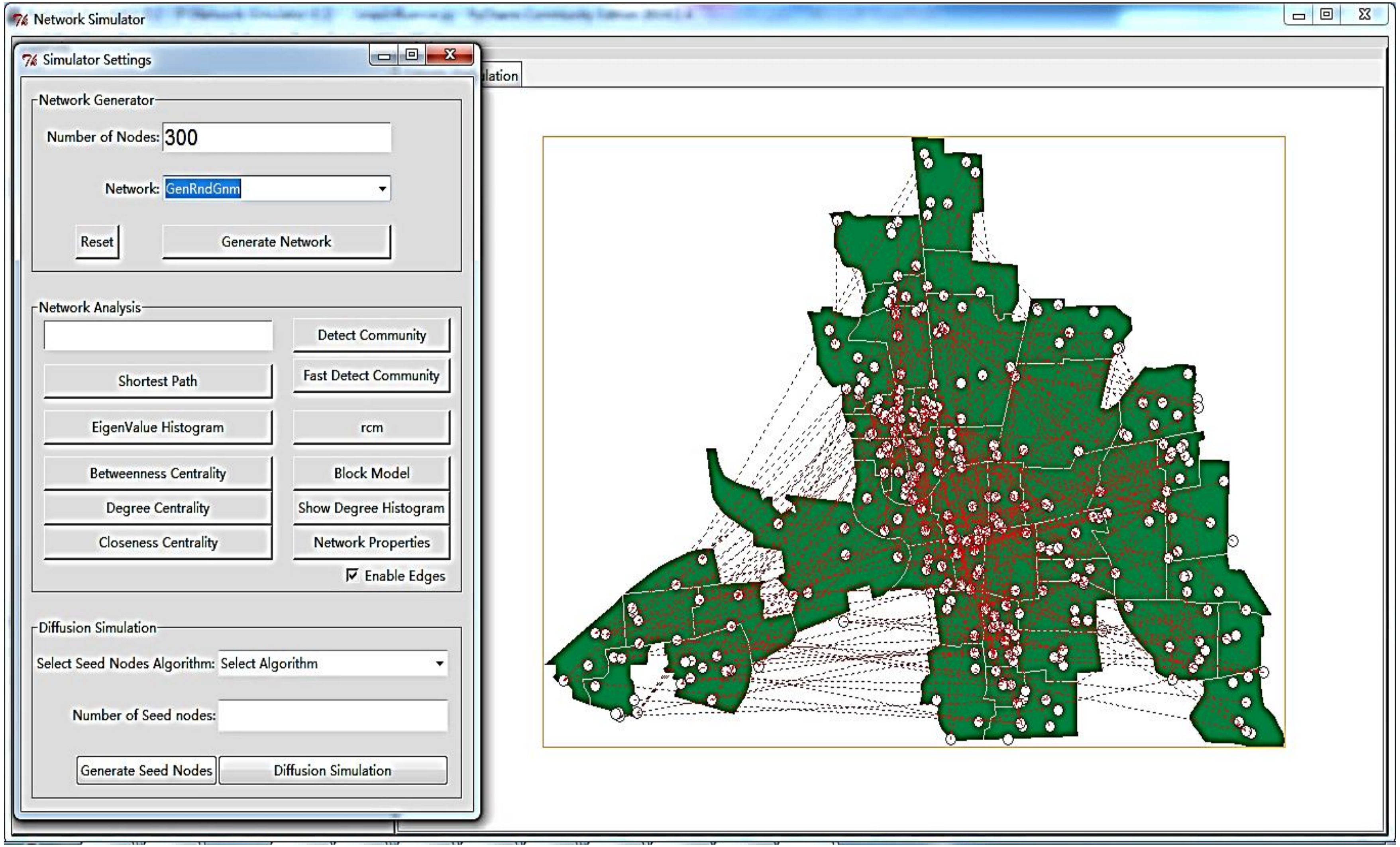
Research Questions

- To what extent the structure of a social network, for example, the different classic network structures, facilitate the process of information diffusion?
- To what extent would social networks account for the process of information diffusion since information does not always spread through social links, i.e., other avenues being the traditional channels of TV/radio/newspaper broadcasting? Use real network data.
- How many early adopters (seed nodes) would be needed to disseminate the information in a certain social network so to ensure wide enough coverage and where are their best locations in the network if to achieve such (the identification methods of early adopters)?

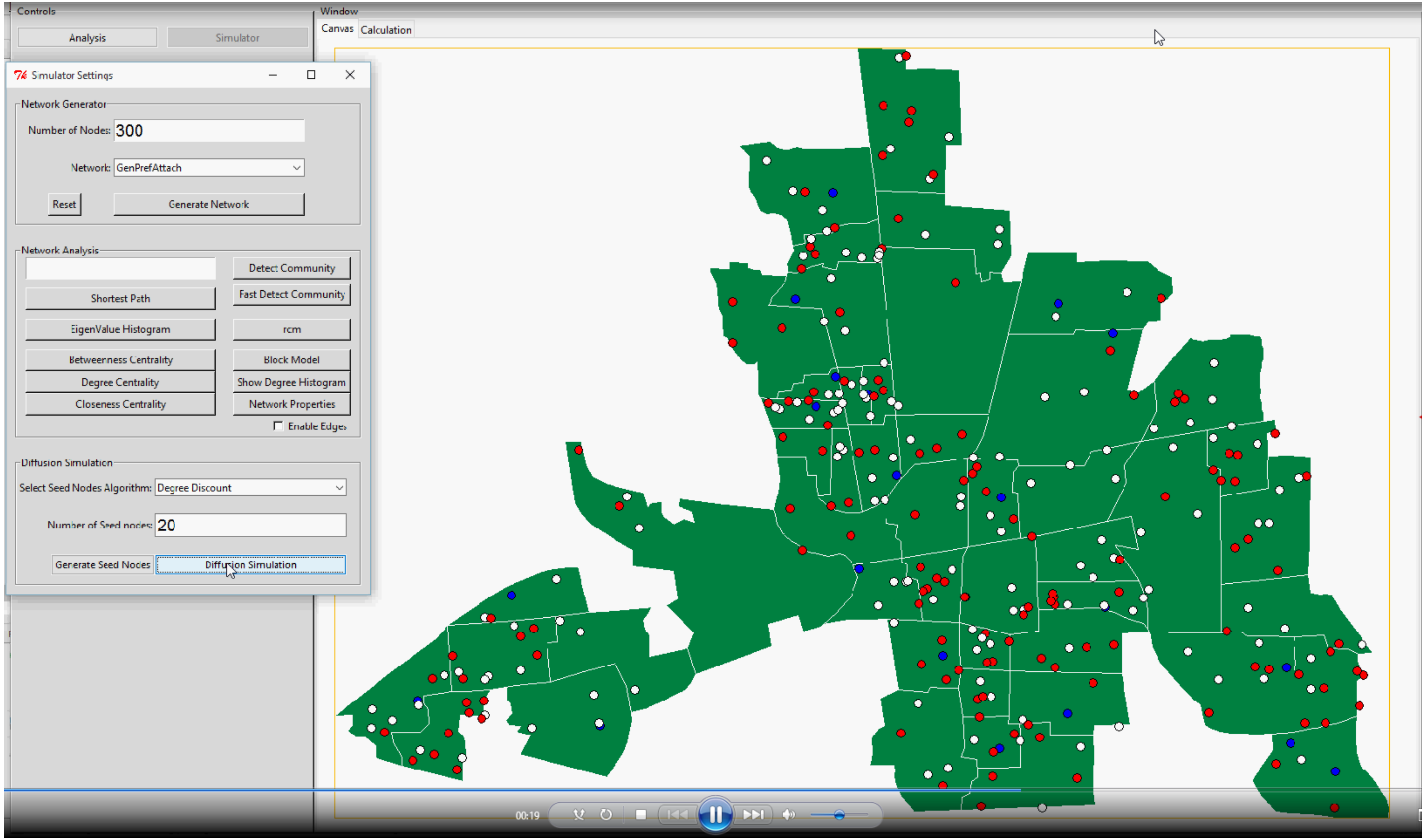
Software Architecture



Network Generator



Simulating Diffusion



Network Analysis

Network analysis was designed to explore the characteristics of a network.

Statistic characteristics of networks

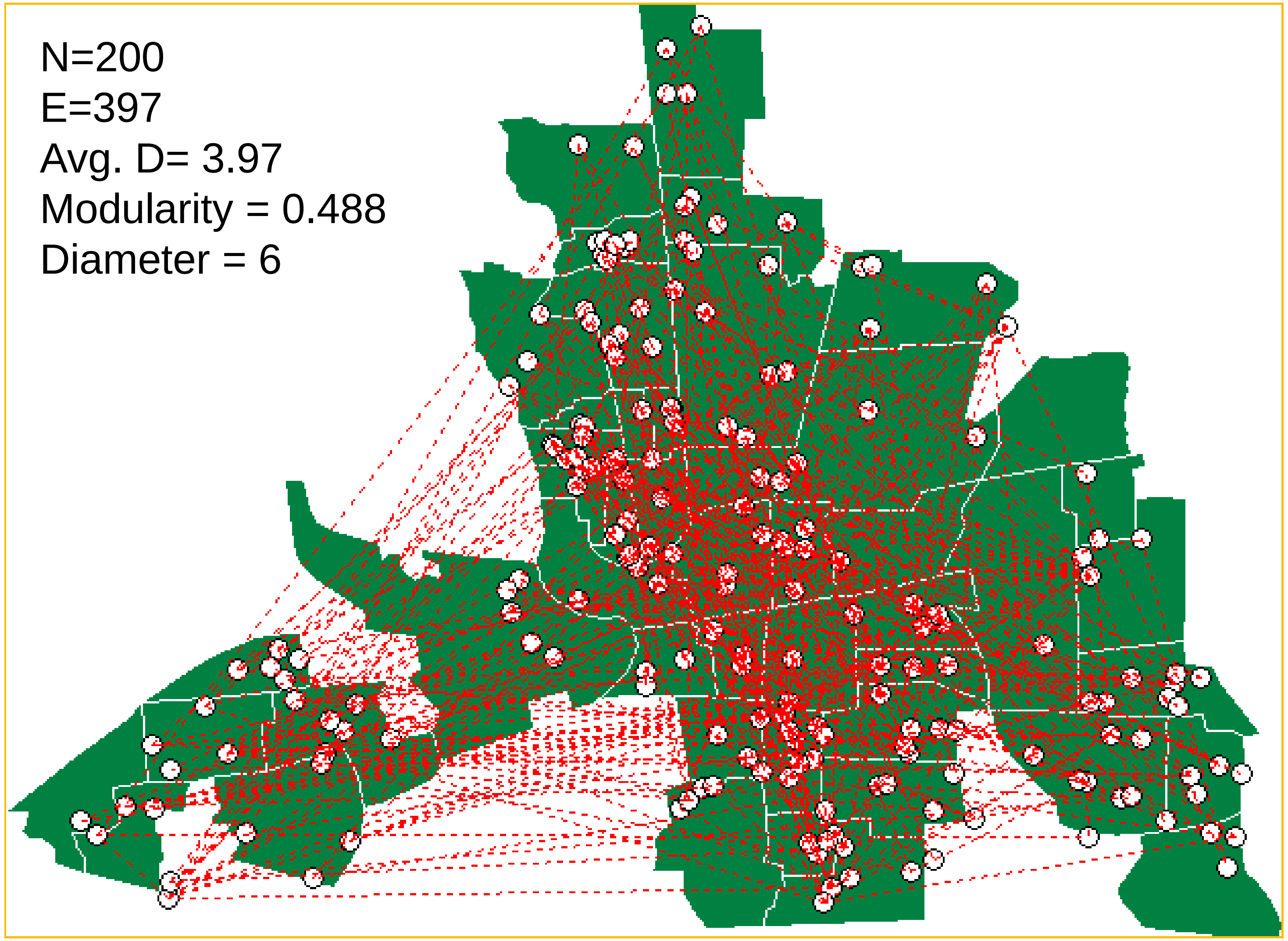
- Number of nodes
- Number of edges
- Modularity, and
- Diameter

Centrality (nodes):

- Degree
- Betweenness
- Closeness
- Eigenvector

Network Analysis

N=200
E=397
Avg. D= 3.97
Modularity = 0.488
Diameter = 6



```
Degree Centrality
---------------------------
0: 0.0502512562814
1: 0.140703517588
2: 0.0954773869347
3: 0.130653266332
4: 0.115577889447
5: 0.115577889447
6: 0.0402010050251
7: 0.0502512562814
8: 0.035175879397
9: 0.0402010050251
10: 0.0804020100503
11: 0.0402010050251
12: 0.0150753768844
13: 0.0301507537688
14: 0.0100502512563
15: 0.0753768844221
16: 0.0452261306533
17: 0.035175879397
18: 0.0251256281407
19: 0.0251256281407
20: 0.0452261306533
21: 0.0100502512563
22: 0.0804020100503
23: 0.0201005025126
```

```
--------------------------------
Betweenness Centrality
--------------------------------
0: 0.0469314647296
1: 0.231325451293
2: 0.115469049112
3: 0.192778644315
4: 0.169508022367
5: 0.16661140247
6: 0.0291526194093
7: 0.053529079177
8: 0.0242713253976
9: 0.0331159380305
10: 0.106586233165
11: 0.0281273348762
12: 0.00615982922345
13: 0.0140235180158
14: 0.0
15: 0.0802420619209
16: 0.0408266635466
17: 0.0213548391579
18: 0.0114090523635
19: 0.0121141411781
20: 0.0387780484049
21: 0.00138947254617
22: 0.0903072755418
23: 0.00858366714922
24: 0.0
```

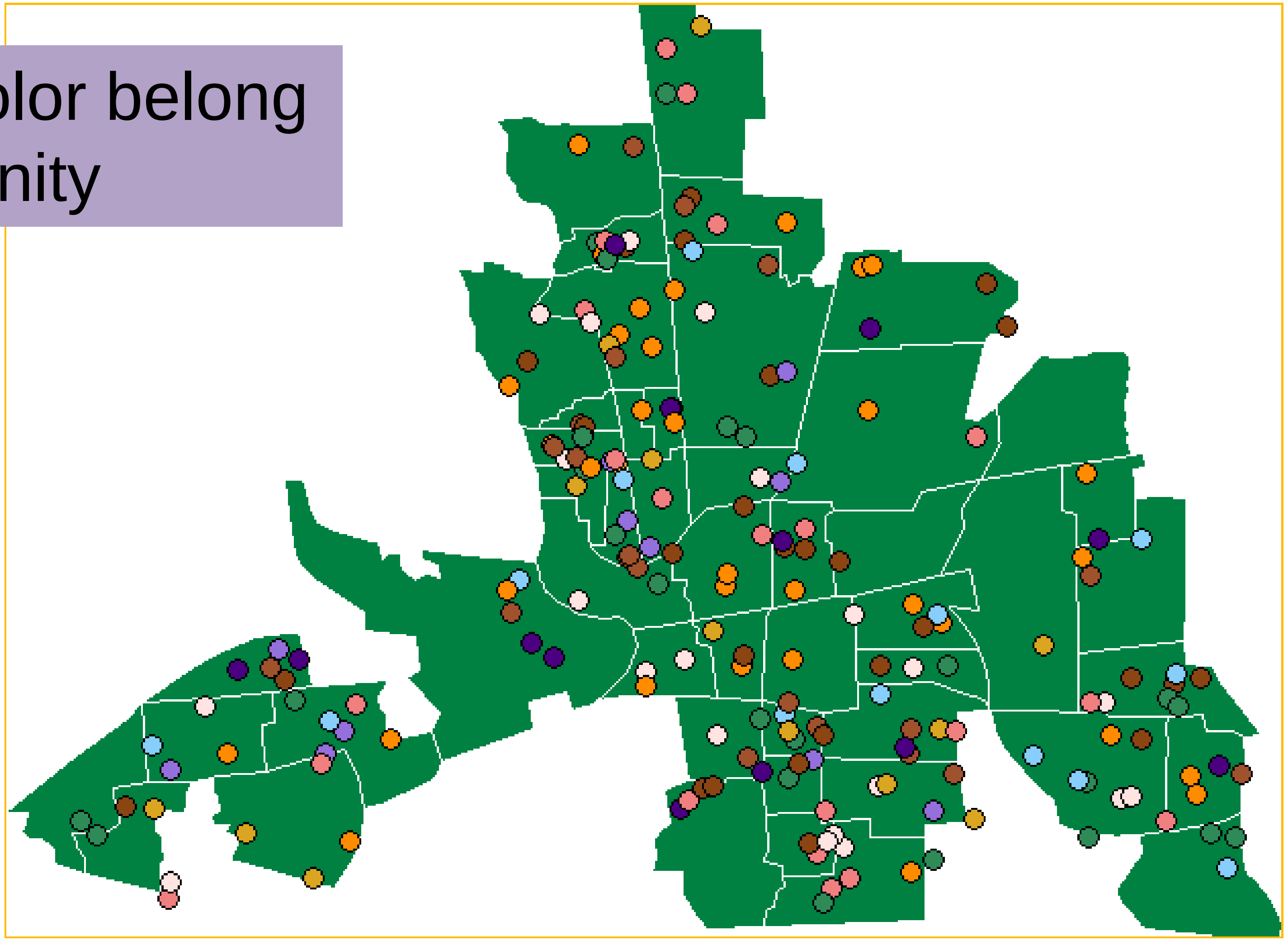
Community Detection

Three methods:

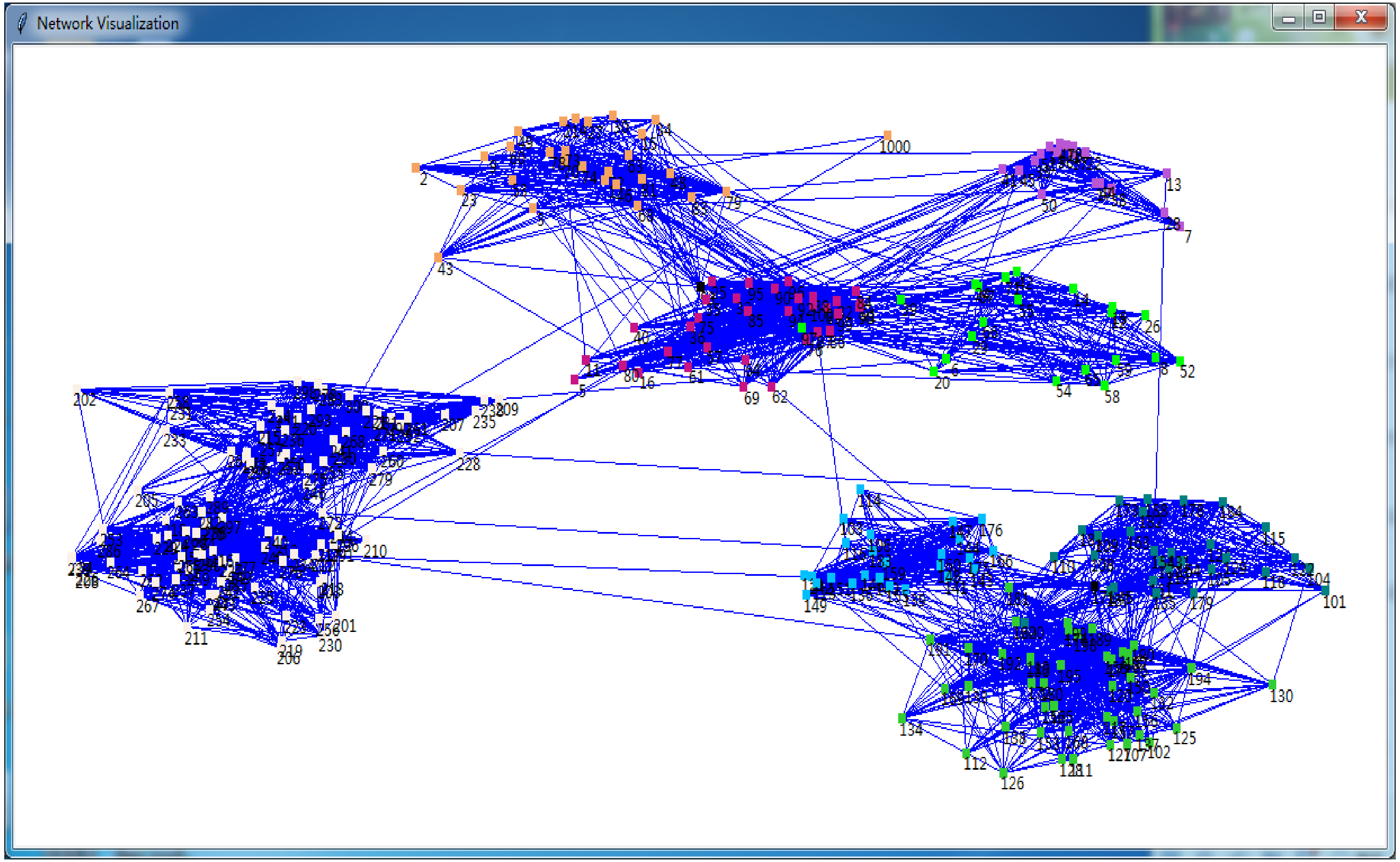
1. Order Statistics Local Optimization method
   Statistical significance, directed/undirected, with/without weights
2. Clauset-Newman-Moore community detection
   Very large networks
3. Girvan-Newman community detection
   Based on betweenness, progressively removing links until left with those between communities

Community Detection

Nodes with same color belong to the same community



Community Visualization



Information Diffusion Simulation

Six centralities and heuristic algorithms for selecting seed nodes:

1. Greedy algorithm
2. Degree discount
3. K-shell
4. Betweenness centrality
5. Closeness centrality
6. Eigenvector centrality

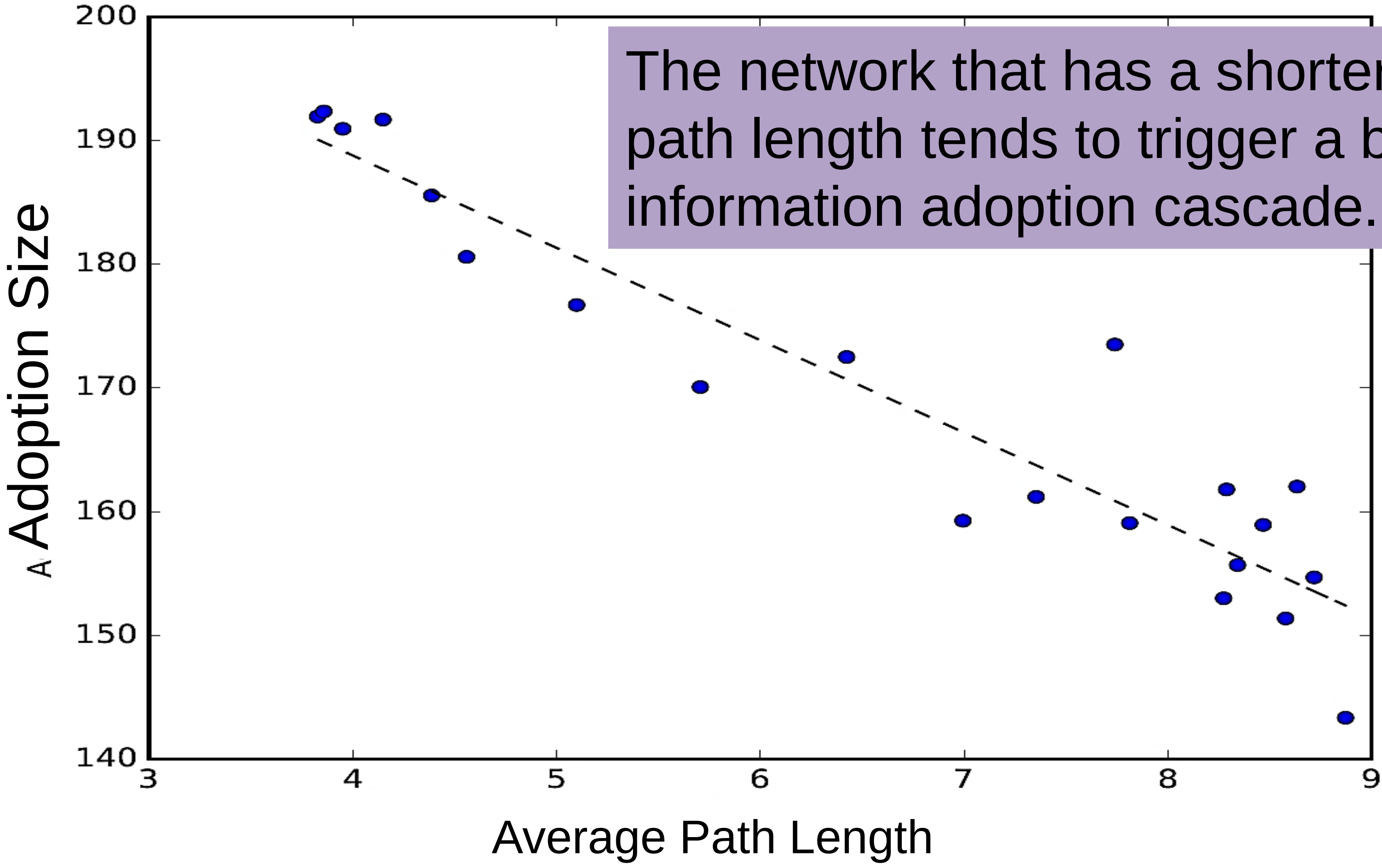
Information Diffusion Simulation

Two information diffusion models:

1. Linear threshold model
2. Independent cascade model

Analyses and Results - I

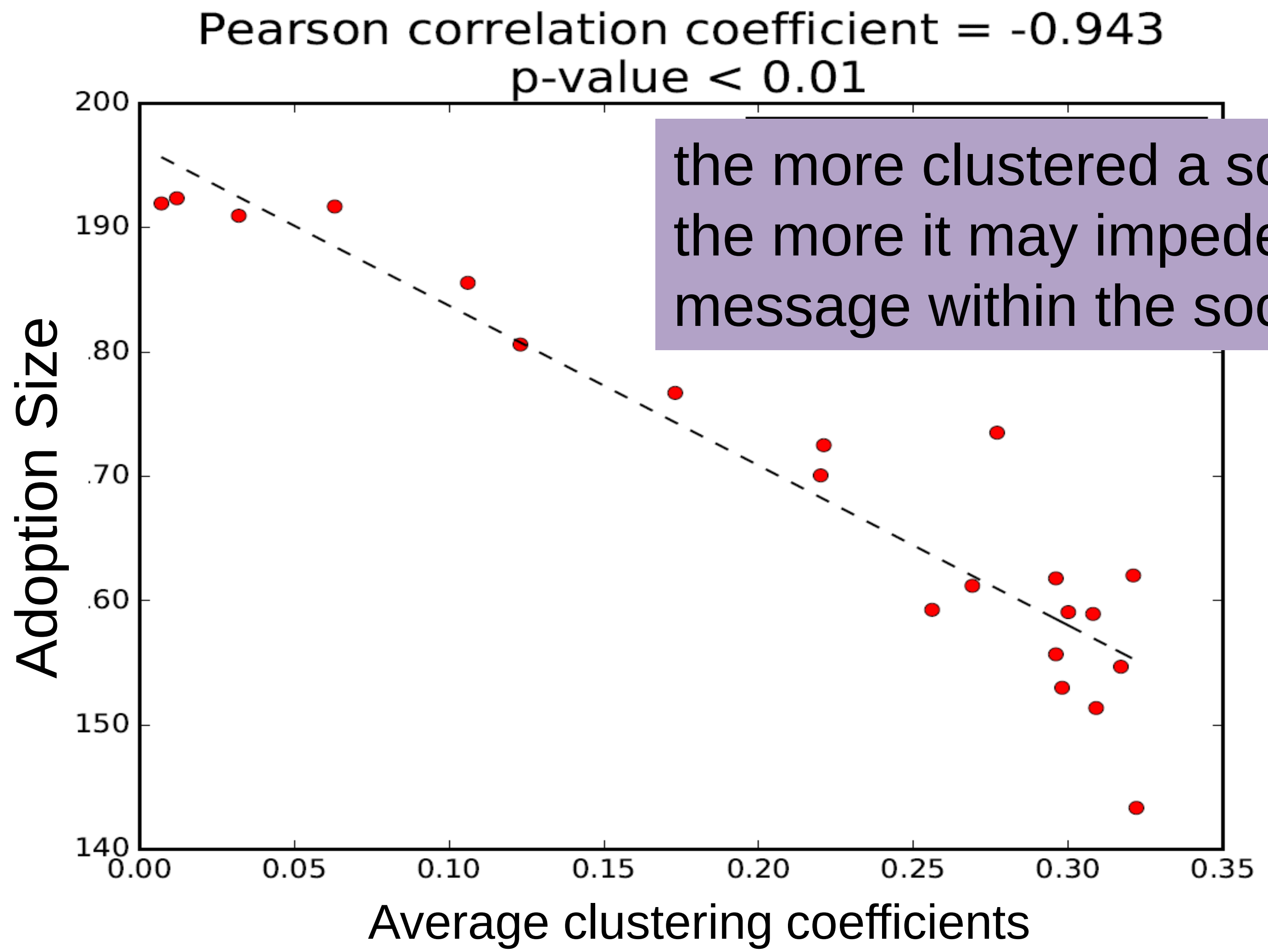
Network Topology and Spread Efficiency



The network that has a shorter average path length tends to trigger a bigger information adoption cascade.

Adoption Rate vs. Avg Clustering Coeff.

Scatter plot of average clustering coefficients c and adoption size f.



the more clustered a social network is, the more it may impede the spread of message within the social network

Analyses and Results - II

Centralities and Heuristics Experiments

- Six different centralities and heuristics were tested on the type of structure in the networks: ***preferential attachment, random, small-world, and lattice networks***, with different number of nodes (*N = 800, 400, and 200*)

- Each type of networks was used in the simulations with three different sets of propagation probabilities:

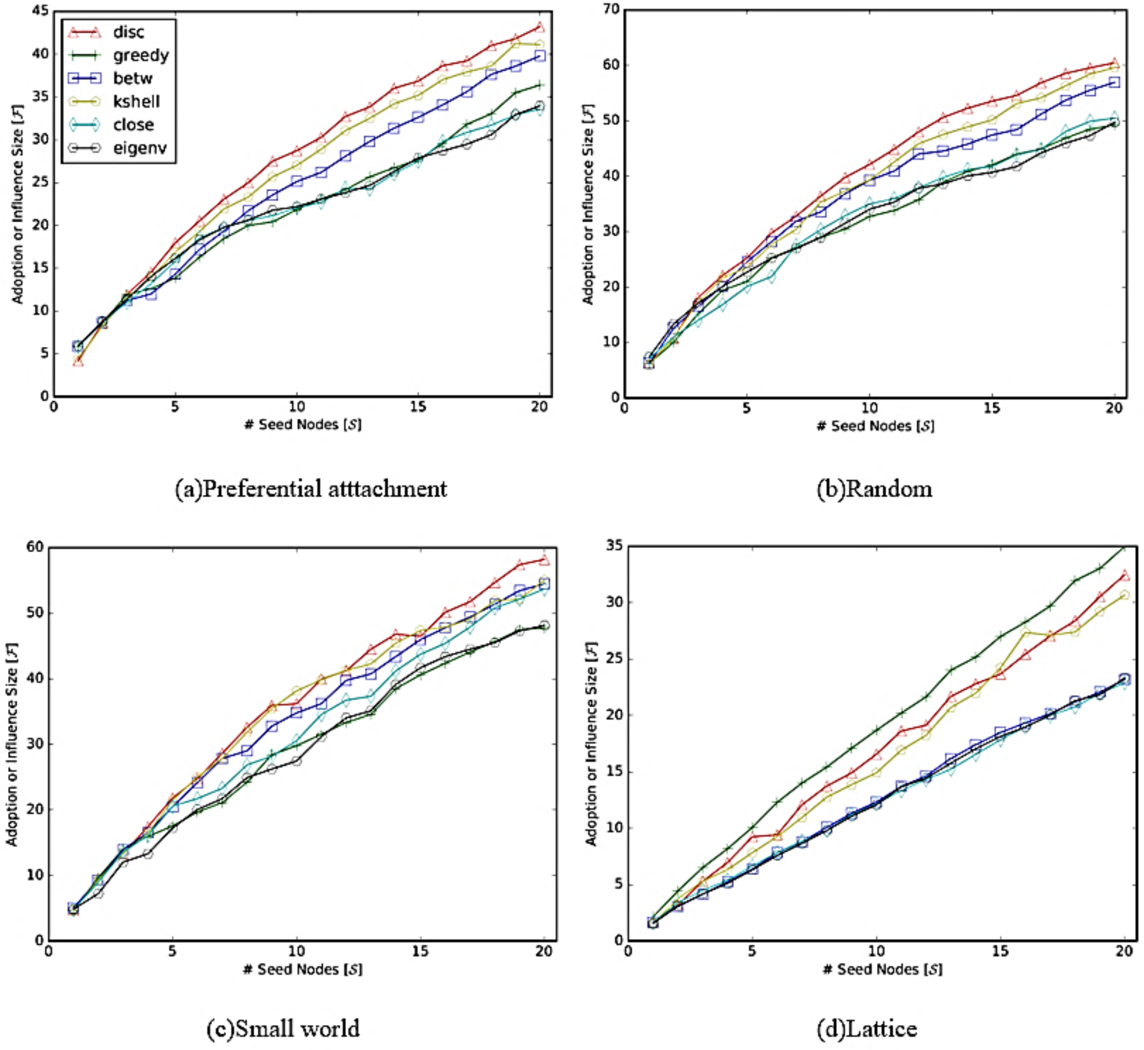
  a) p_{op} =0.4, p_n =0.3,

  b) p_{op}=0.3, p_n=0.2 and

  c) p_{op}=0.2, p_n=0.1.

Analyses and Results - Continued

- Information diffusion on **four types** of simulated networks with **six** centralities and heuristics.
- The total number of nodes is $N = 200$; Propagation probabilities for opinion leaders and normal people are $\boldsymbol{p_{op} = 0.2, p_n = 0.1}$.

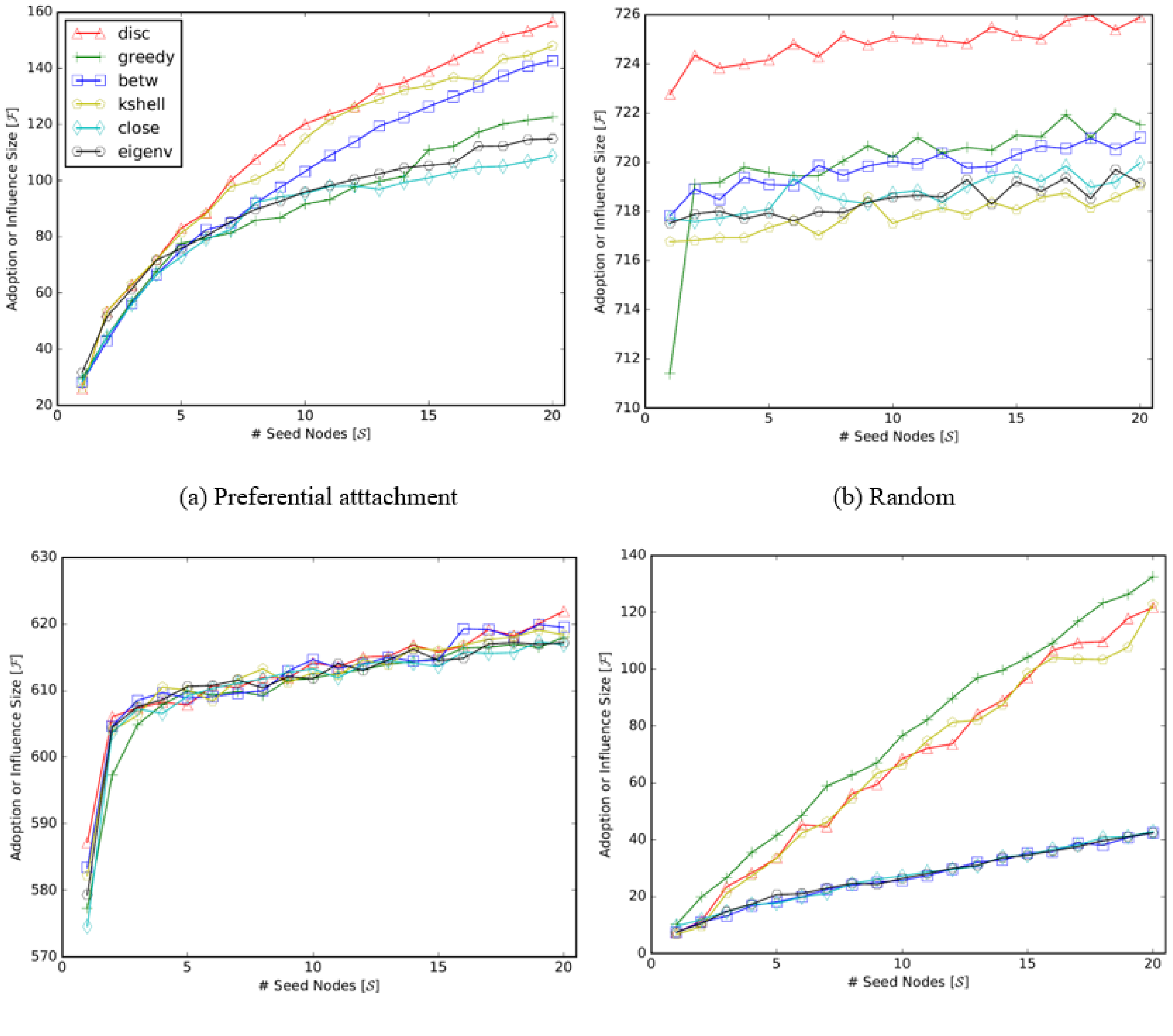


Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality

More seed nodes → Higher adoption size

Analyses and Results - Continued

- Information diffusion on four types of artificial networks with six centralities and heuristics.
- The total number of nodes is $N = \mathbf{800}$; Propagation probabilities for opinion leaders and normal people are $\boldsymbol{p_{op}} = \mathbf{0.4}, \boldsymbol{p_n} = \mathbf{0.3}$



(a) Preferential atttachment

(b) Random

(c) Small world

(d) Lattice

For larger networks, seed nodes are even more critical

Analyses and Results - III

- Evaluation of the **efficiency of diffusion** on each early adopter in three different networks.
- S denotes the **rank of seed nodes** (early adopters); PA refers to **preferential attachment**. The network size $N = 200$

	p_{op}=0.2			p_{op}=0.3			p_{op}=0.4		
S	PA	small-world	random	PA	small-world	random	PA	small-world	random
1	3.204	3.7	5.316	7.427	27.155	37.945	11.554	133.455	121.207
2	3.384	3.847	3.471	4.054	13.608	9.761	10.045	8.394	9.432
3	2.351	2.77	6.314	4.895	7.552	12.217	3.455	0.581	-0.878
4	1.597	3.122	3.014	3.634	10.439	4.619	5.247	1.326	0.447
5	2.441	3.396	2.113	2.065	6.141	2.394	3.201	-0.199	0.46
6	1.576	1.935	3.532	2.708	6.36	3.202	3.021	0.476	0.352
7	1.519	2.868	2.032	2.059	2.15	3.651	2.421	0.384	-1.505
8	0.907	2.965	2.614	1.959	3.263	1.453	0.049	-0.531	0.295
9	1.524	2.355	2.386	-0.193	2.503	2.367	5.168	-0.57	0.526
10	0.235	-0.757	1.337	3.229	-0.601	0.987	1.706	-0.185	-1.54
11	0.546	2.776	1.742	0.791	1.717	1.335	1.628	-0.458	0.355
12	1.443	0.23	2.121	0.745	-0.517	0.992	0.698	-0.547	-0.551
13	0.113	2.315	1.602	0.877	4.669	1.035	0.433	-0.516	-0.355
14	1.183	1.298	0.675	0.38	0.463	-1.547	0.595	-0.042	-0.384
15	-0.165	-1.338	0.271	1.267	-1.206	2.416	0.917	-0.662	-0.904
16	0.797	2.68	0.034	-0.153	0.486	0.081	0.73	0.011	-0.414
17	-0.43	0.617	1.285	0.669	1.698	1.02	1.073	-1.095	-0.603
18	0.783	1.929	0.689	0.722	1.132	-0.669	-0.451	-0.203	-0.463
19	-0.191	1.717	0.001	-0.205	0.447	-0.843	0.915	-1.263	-0.903
20	0.389	-0.194	-0.085	0.361	2.158	1.591	-0.279	-0.166	-0.14

Propagation probability affects small-world networks and random networks the most

Analyses and Results – IV

- Real Network Examination
  - Bernardo wildfire tweets:

Table. Result of information diffusion in Bernardo wildfire tweets under 5-day partition.

Time-range	Seed accounts	Influence
Day 1	KUSI_News, RSF_Fire	622
Day 2	SanDiegoCP, thesandiegonewz, twit_san_diego, sandiegobnews, 10News, ooph, dancohenCBS8	447
Day 3	blufinki	142
Day 4	BlazonLaurels, EdZieralski,jennifercdougla	213
Day 5	thesandiegonewz, AthensMarketSD, KPBSnews	52

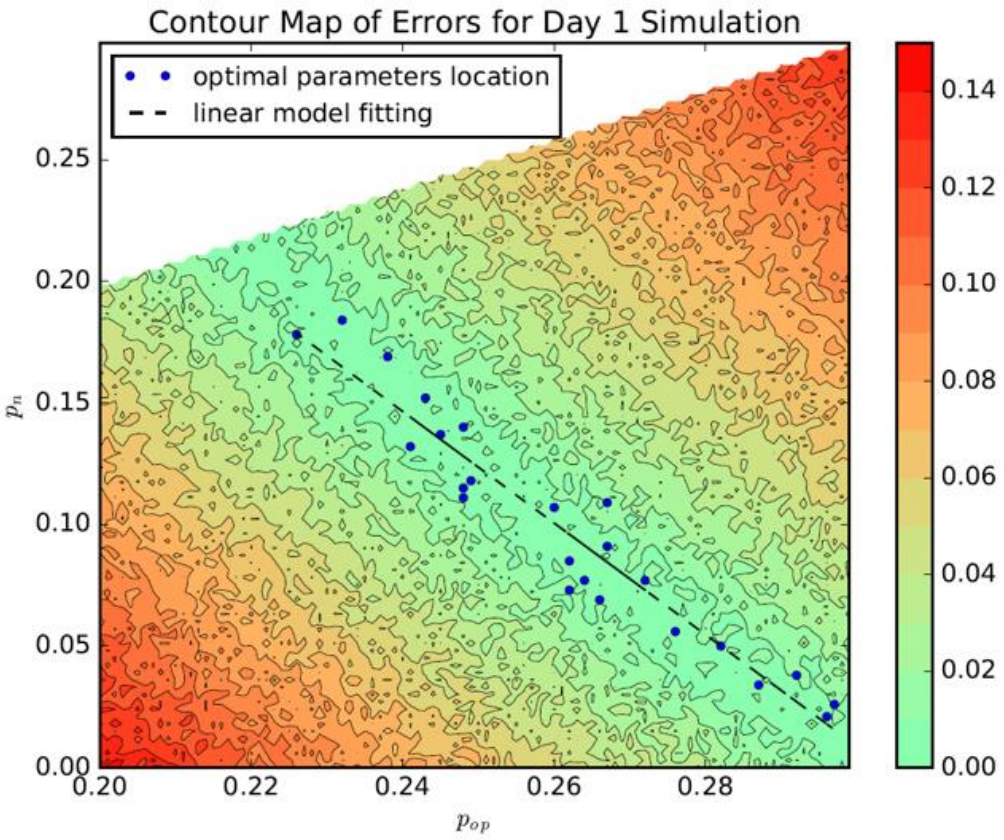
Analyses and Results – V

❖ **Real Network** Examination

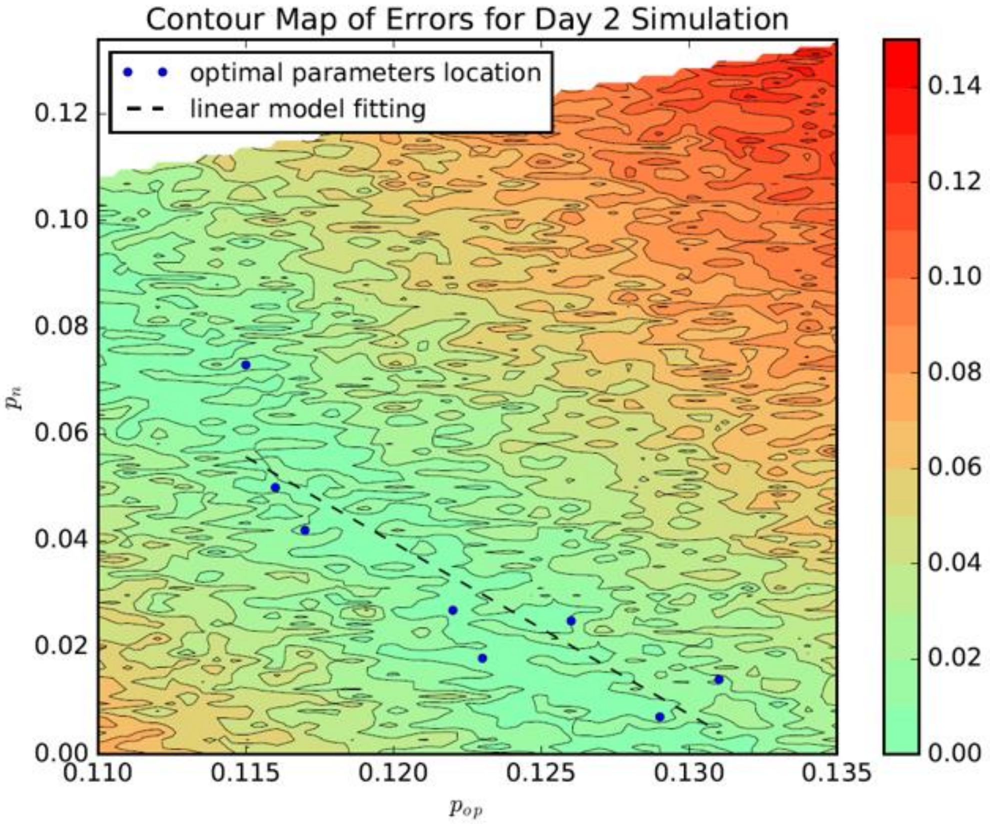
- To find the parameters of the information diffusion model that could mimic the information diffusion in Bernardo wildfire:
  - Bernardo wildfire network was imported
  - Grid search was conducted for emulating different propagation probabilities
    - With a fixed increment of 0.001, starting with 0.1 in each simulation.

Analyses and Results – Continued

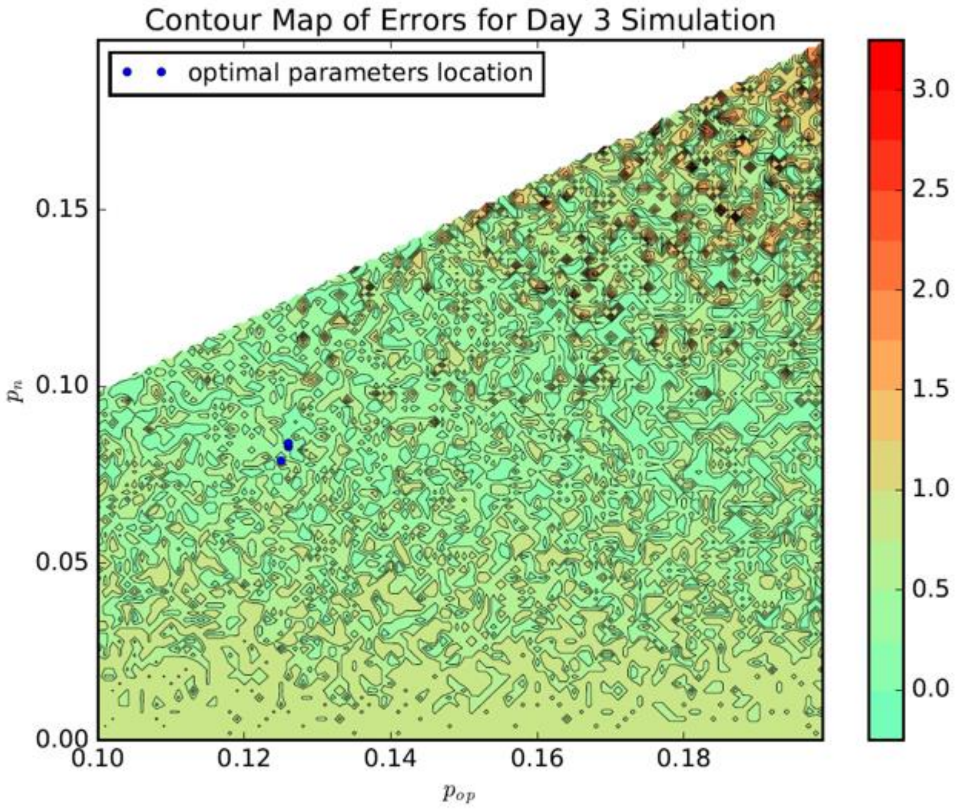
- Using contour maps, the areas of errors (difference between simulated results and real network results) was plotted from low to high.
- Lower values are light greens, and higher values are dark reds.



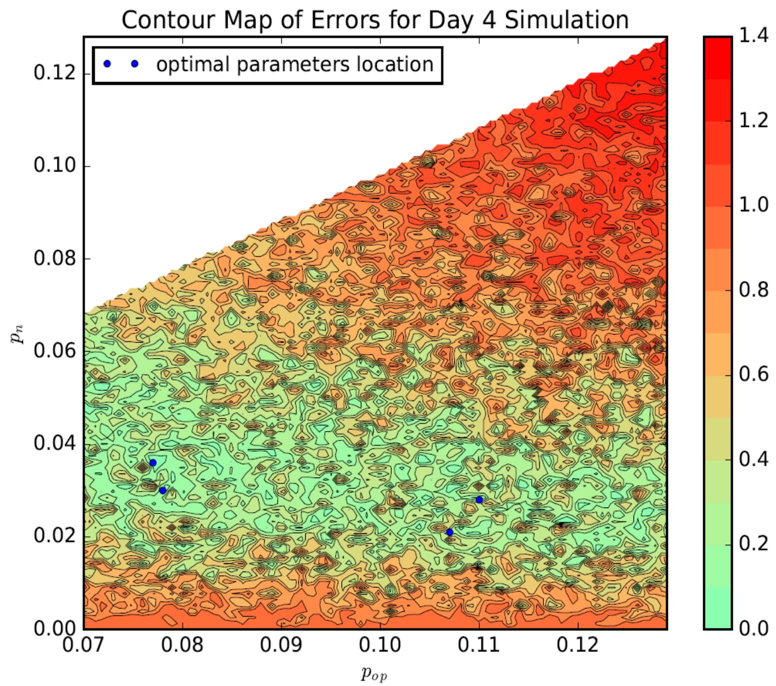
Contour map illustrating the sensitivity of information diffusion model on the **Day 1** of Bernardo wildfire data.



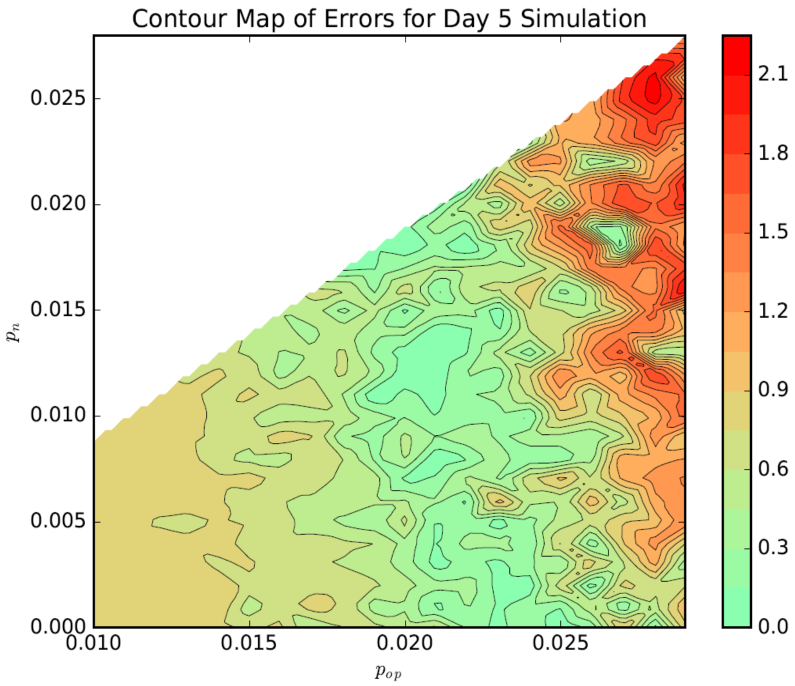
Contour map illustrating the sensitivity of information diffusion model on the **Day 2** of Bernardo wildfire data.



Contour map illustrating the sensitivity of information diffusion model on the **Day 3** of Bernardo wildfire data.



Contour map illustrating the sensitivity of information diffusion model on the **Day 4** of Bernardo wildfire data



Contour map illustrating the sensitivity of information diffusion model on the **Day 5** of Bernardo wildfire data

Real Network Examination Summary

Simulation tools **could** mimic the information diffusion in real events.

- **Range** of optimal parameters (propagation probabilities)
- Parameters of **propagation probabilities** usually decrease along with time unless a new update emerged in the topic.

Concluding Remarks

Efficient Information Diffusion is determined by:

Network structure

A shorter average path length or a lower average clustering coefficient tended to have a wider information diffusion

Influential early adopters

Degree discount performs the best over all types of ntwk

Greedy only performs well in Lattice network

Well-connected networks need fewer early adopters

Propagation probability

Higher propagation probability leads to more efficient information diffusion, needing fewer early adopters

Future Network

- Improve existing tools to be suitable for **large scale networks**
- Develop additional social network **tools**
- Improve **influence maximization** algorithms for better understanding and more effectively predicting the spread in the social network with spatial and temporal content.
- **Spatial clustering** and **Social clustering**