Correlation Analysis of Node Importance Measures

An Empirical Study through Graph Robustness

Stony Brook University
Computer Science

Mirza Basim Baig
Leman Akoglu

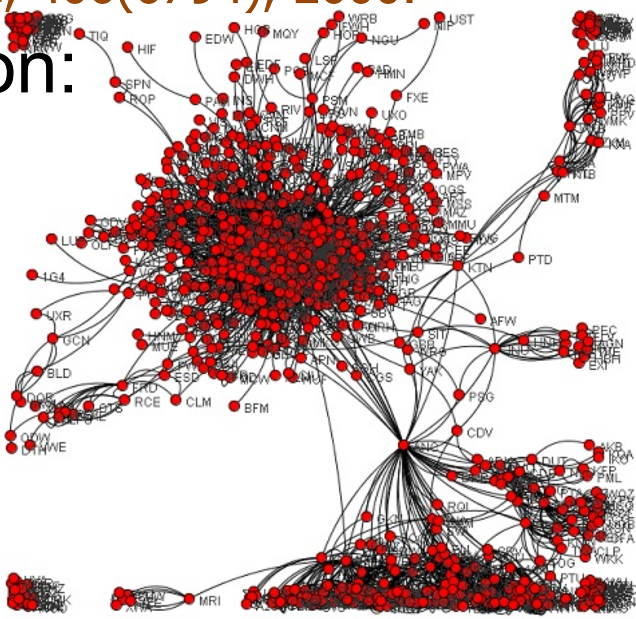
Node Importance

Fundamental in network analysis:

- finding central/influental/core nodes
- measuring attack-tolerance
  - Real graphs are vulnerable to targeted attacks

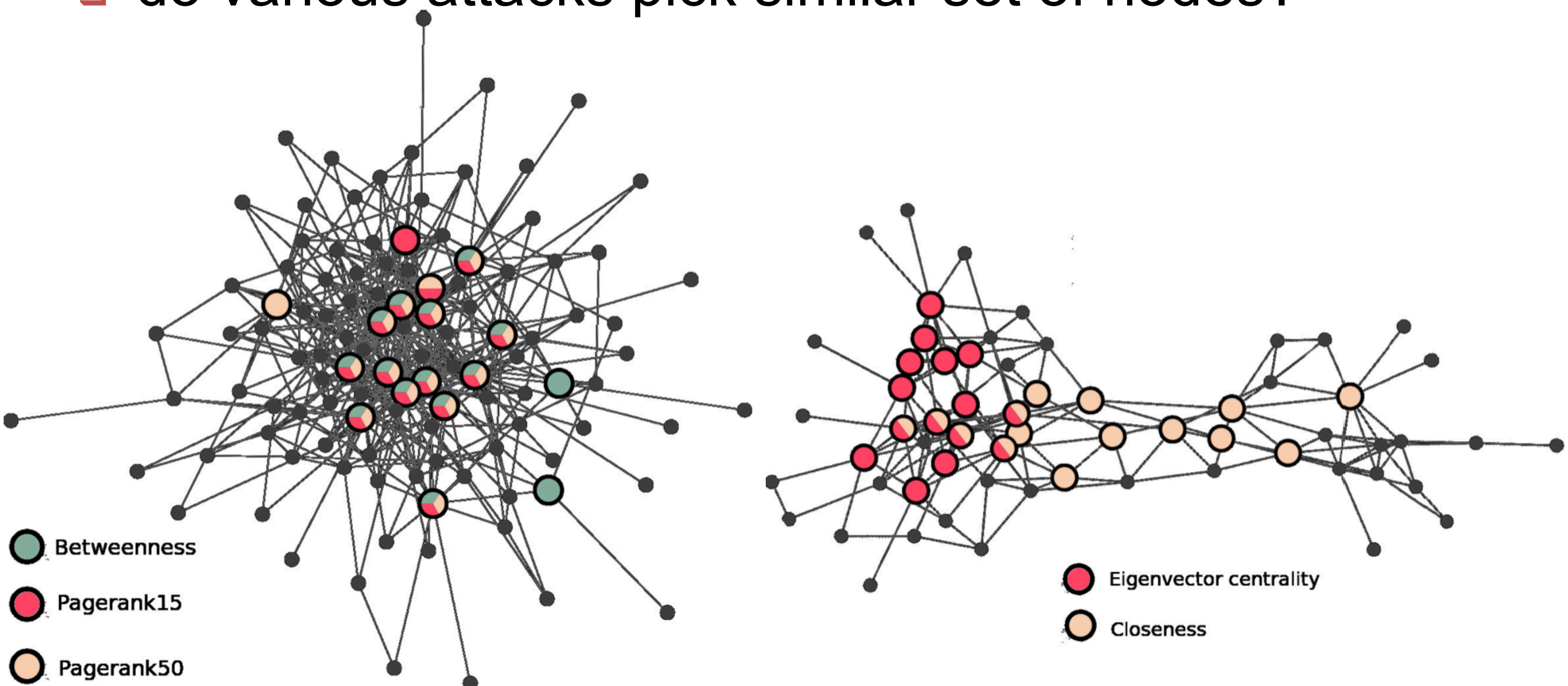
    R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks. Nature, 406(6794), 2000.
  - Numerous strategies, based on:

    Pagerank
    Betweenness
    Closeness
    Katz
    …



Node Importance Measures

- How similar are different importance measures?
  - do various attacks pick similar set of nodes?



(a) Significant overlap among node sets

(b) Small overlap among node sets

http://www-personal.umich.edu/ mejn/netdata/

This Work:

Empirical analysis of correlations between node importance measures

Goal:

- Reduce long list of measures into groups such that
- cheaper alternatives to complex measures
- a few proxies for consensus finding

Related Work

- Correlation analysis of centralities:
  - **Bolland**, 1988 – 4 measures
  - **Rothenberg** et. al, 1995 – 8 measures
  - **Valente** et. al, 2008 – 4 measures
  - → studied very small graphs (hundreds of nodes), from one domain (often social), with single method
  - **Vigna**, 2015 – 5 measures, one large graph
- Correlation of algorithms/measures
  - **Abrahao** et. al, 2012 – clustering algorithms
  - **Soundarajan** et. al, 2014 – graph similarity measures

This Work:

Stony Brook University
Computer Science

Empirical analysis of correlations between node importance measures

- 15 measures (i.e., attack strategies)
  - randomized, local, distance, spectral
- 68 real-world graphs
  - social, bio, infra, info
- 3 analysis approaches
  - I) rankings, II) node types, III) graph disruption
- Analysis results

Node attack strategies (I)

	Id	Abbr.	Description	bigO
Random	1	r	Random node	$O(k)$
	2	rn	Random neighbor of a randomly picked node	$O(k)$
	3	rw10	Most visited node in a random walk of length $T = 10$	$O(kT)$
	4	rw50	Most visited node in a random walk of length $T = 50$	$O(kT)$

Node attack strategies (II)

	Id	Abbr.	Description	bigO
Local	5	deg	Highest degree	$O(m)$
	6	lcc	Highest local clust. co-efficient [38]	$O(nd^3)$
	7	ecc	Highest extended clustering co-efficient [12]	$O(nd^{2+D})$

Node attack strategies (III)

	Id	Abbr.	Description	bigO
Dist.	8	rad	Lowest radius [13]	$O(n^3)$
	9	cc	Highest closeness centrality [26]	$O(n^3)$
	10	betw	Highest betweenness centrality [5]	$O(nm)$

Node attack strategies (IV)

Stony Brook University
Computer Science

	Id	Abbr.	Description	bigO
Spectral	11	eig	Highest eigen-vector centrality	$O(nC)$
	12	pr15	Highest PageRank [27] (α=0.15)	$O(mt)$
	13	pr50	Highest PageRank [27] (α=0.50)	$O(mt)$
	14	katz	Highest Katz index [17]	$O(mt)$
	15	comm	Highest self-communicability [10]	$O(n^3)$

15 node attack strategies

	Id	Abbr.	Description	bigO
Random	1	r	Random node	$O(k)$
	2	rn	Random neighbor of a randomly picked node	$O(k)$
	3	rw10	Most visited node in a random walk of length $T = 10$	$O(kT)$
	4	rw50	Most visited node in a random walk of length $T = 50$	$O(kT)$
Local	5	deg	Highest degree	$O(m)$
	6	lcc	Highest local clust. co-efficient [38]	$O(nd^3)$
	7	ecc	Highest extended clustering co-efficient [12]	$O(nd^{2+D})$
Dist.	8	rad	Lowest radius [13]	$O(n^3)$
	9	cc	Highest closeness centrality [26]	$O(n^3)$
	10	betw	Highest betweenness centrality [5]	$O(nm)$
Spectral	11	eig	Highest eigen-vector centrality	$O(nC)$
	12	pr15	Highest PageRank [27] (α=0.15)	$O(mt)$
	13	pr50	Highest PageRank [27] (α=0.50)	$O(mt)$
	14	katz	Highest Katz index [17]	$O(mt)$
	15	comm	Highest self-communicability [10]	$O(n^3)$

This Work:

Stony Brook University
Computer Science

Empirical analysis of correlations between node importance measures

- 15 measures
  randomized, local, distance, spectral
- 68 real-world graphs
  social, bio, infra, info
- 3 analysis approaches
  I) rankings, II) node types, III) graph disruption
- Analysis results

Real-world graphs

Stony Brook University
Computer Science

Id	Name	Type	Nodes	Edges
1	bo-bio	Bio	1458	5393
2	clegans	Bio	453	4394
3	mintcaenor	Bio	3026	13484
4	mintmammals	Bio	7836	41356
5	mintvirus	Bio	950	3478
6	pollen1	Bio	793	6638
7	pollen2	Bio	766	3133
8	pollen3	Bio	712	2917
9	pollen4	Bio	997	4810
10	seed-dispersion1	Bio	209	1521
11	seed-dispersion2	Bio	317	2527
12	yeasts	Bio	2224	15874
13	coauth2	Information	21363	203989
14	coauth3	Information	4158	31003
15	coauth5	Information	8638	58229
16	csphd	Information	1025	3110
17	jazz	Information	198	5661
18	pgp	Information	10680	59310

* 12 biological networks
* 6 information networks

Real-world graphs

19	caida6-1	Infra	21202	107050
20	caida6-2	Infra	21157	106623
21	caida6-3	Infra	21232	106974
22	caida6-4	Infra	21245	105770
23	caida6-5	Infra	21339	107459
24	caida7-1	Infra	24013	122185
25	caida7-2	Infra	24018	121913
26	caida7-3	Infra	24056	122342
27	caida7-4	Infra	24078	121700
28	caida7-5	Infra	20906	106460
29	oregon1-1	Infra	10670	54646
30	oregon1-2	Infra	10729	54698
31	oregon1-3	Infra	10790	55164
32	oregon1-4	Infra	10859	55780
33	oregon1-5	Infra	10886	55300
34	oregon1-6	Infra	10943	55590

* 36 infrast. networks

Real-world graphs

35	oregon1-7	Infra	11011	55798
36	oregon1-8	Infra	11051	55937
37	oregon1-9	Infra	11174	57427
38	oregon2-1	Infra	10900	73225
39	oregon2-2	Infra	10981	72656
40	oregon2-3	Infra	11019	74505
41	oregon2-4	Infra	11080	74118
42	oregon2-5	Infra	11113	73945
43	oregon2-6	Infra	11157	73007
44	oregon2-7	Infra	11260	73830
45	oregon2-8	Infra	11375	75910
46	oregon2-9	Infra	11461	76881
47	p2p4	Infra	10876	90850
48	p2p5	Infra	8842	72498
49	p2p6	Infra	8717	71763
50	p2p8	Infra	6299	47850
51	p2p9	Infra	8104	60111
52	p2p24	Infra	26498	157215
53	p2p25	Infra	22663	132047
54	p2p30	Infra	36646	213246

Real-world graphs

55	california-cell	Social	1718	9743
56	egoFacebook	Social	2888	8849
57	enron	Social	33696	395248
58	emailURV	Social	1133	12013
59	pennsylvania-cell	Social	2514	14391
60	wiki	Social	7066	208509
61	slashdot	Social	77360	1015534
62	anybeat	Social	12645	106109
63	small-company1	Social	320	5042
64	small-company2	Social	165	1609
65	medium-company1	Social	1429	40014
66	medium-company2	Social	3862	178470
67	large-company1	Social	5793	67298
68	large-copmany2	Social	5524	193906

* 14 social networks

All datasets available at: https://github.com/basimbaig/robust14

This Work:

Empirical analysis of correlations between node importance measures

- 15 measures

  randomized, local, distance, spectral
- 68 real-world graphs

  social, bio, infra, info
- 3 analysis approaches

  I) rankings, II) node types, III) graph disruption
- Analysis results

Meta-approach

Given strategies 1…M=15, set of (68) graphs G

- Compute similarity s_{ij} between all pairs i, j
- Construct MxM similarity matrix S
- Hierarchically cluster (complete-linkage) S

Output clusters in majority (>50%) of G

Meta-approach

Stony Brook University
Computer Science

Given strategies 1…M=15, set of (68) graphs G

- Compute **similarity s_{ij}** between all pairs i, j
- Construct MxM similarity matrix S
- Hierarchically cluster (complete-linkage) S

Output clusters in majority (>50%) of G

3 approaches to similarity:

I.	RANK-C	ranking
II.	Top$k$-C	node characteristics
III.	Response-C	disruption dynamics

Approach I: RANK-C

Stony Brook University
Computer Science

- Compares strategies based on their ranking of all the nodes

1. Rank nodes (ranklist r^i for strategy i)
   - non-randomized: sorted by measure
   - randomized: order of nodes picked
2. Rank correlation by Weighted-Tau [Vigna, 2015]: generalizes Kendall's Tau:
   - ties carefully accounted for
   - correlation biased toward agreement in higher ranks

$$\boxed{S_{ij} = \text{Weighted-Tau}(r^i, r^j) \in [-1,1]}$$

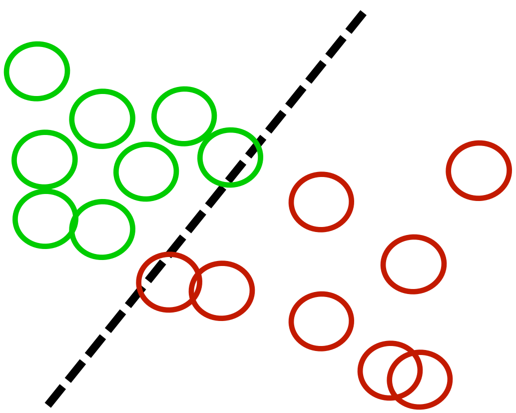
Approach II: Top$k$-C$^{\text{SVM-Sep}}$

Stony Brook University
Computer Science

- Compares strategies based on the *kind* (characteristics) of nodes they select

1. Find top-k nodes for strategies 1…M
2. Extract recursive structural features

   [Hendersen+ 2011]: node → feature vector
3. SVM classifier for k vectors from **i** and **j**



feature vectors
- ○ nodes by strategy i
- ○ nodes by strategy j

Approach II: Top$k$-C$^{\text{SVM-Sep}}$

- Compares strategies based on the *kind* (characteristics) of nodes they select

1. Find top-k nodes for strategies 1…M
2. Extract recursive structural features

   [Hendersen+ 2011]: node → feature vector
3. SVM classifier for k vectors from **i** and **j**

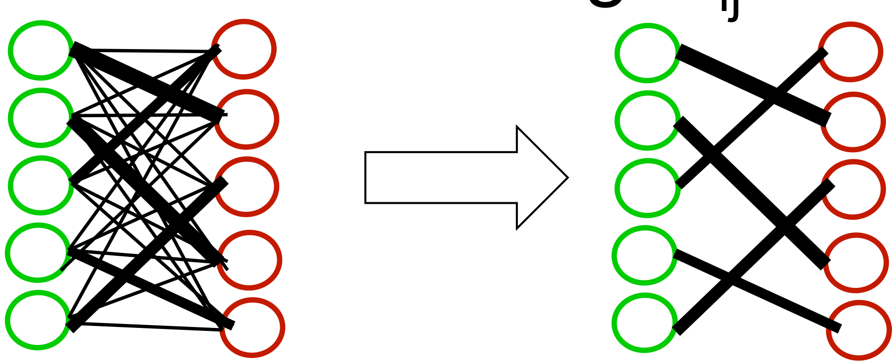
$$\boxed{S_{ij} = 1 - \text{Class-Separability}(V_k^i, V_k^j) \in [0,1]}$$

Class-separability: avg. probability mass of correctly classified node-vectors

Approach II: Top k-C^BI-MATCH^

- Compares strategies based on the *kind* (characteristics) of nodes they select

1. Find top-k nodes for strategies 1…M
2. Extract recursive structural feature vectors
3. Construct complete V_k^i x V_k^j bipartite graph
   - edge weight = vector similarity
4. Find maximum matching m_{ij}^*:



Approach II: Top$k$-C$^{\text{BI-MATCH}}$

Stony Brook University
Computer Science

- Compares strategies based on the *kind* (characteristics) of nodes they select

1. Find top-k nodes for strategies 1…M
2. Extract recursive structural features

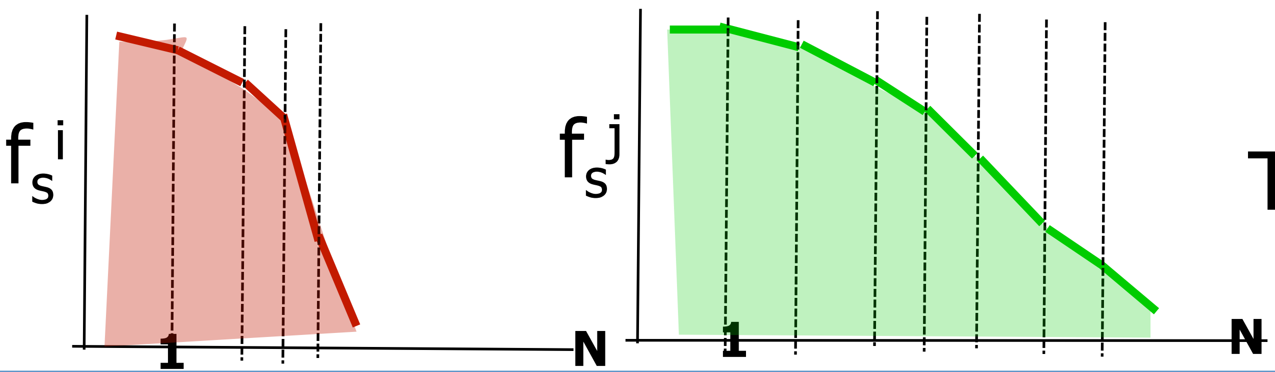
   [Hendersen+ 2011]: node → feature vector
3. Construct complete V_k^i x V_k^j bipartite graph
   - edge weight = vector similarity
4. Find maximum matching m_{ij}^*

$$\boxed{S_{ij} = \text{Total-weight}(m_{ij}^*) / k \in [0,1]}$$

Approach III: Response-C

- Compares strategies based on disruption dynamics they cause when nodes removed

1. Rank nodes for strategy i
2. Remove nodes 1-by-1 in rank order
3. Compute robustness f_s^i when s nodes removed; 1) f=GCC fraction, 2) $f=\lambda_1$
4. Attack-tolerance of g: $T(g)^i = avg(f_s^i)$, s=1…N



$T(g)^i < T(g)^j$

Approach III: Response-C

Stony Brook University
Computer Science

- Compares strategies based on disruption dynamics they cause when nodes removed

1. Rank nodes for strategy i
2. Remove nodes 1-by-1 in rank order
3. Compute robustness f_s^i when s nodes removed; 1) f=GCC fraction, 2) $f=\lambda_1$
4. Attack-tolerance of g: $T(g)^i = avg(f_s^i)$, s=1…N
5. Rank graphs g in G by $T(g)^i$ into R^i

$$\boxed{S_{ij} = \text{Weighted-Tau}(R^i, R^j) \in [-1,1]}$$

This Work:

Empirical analysis of correlations between node importance measures

- 15 measures

  randomized, local, distance, spectral

- 68 real-world graphs

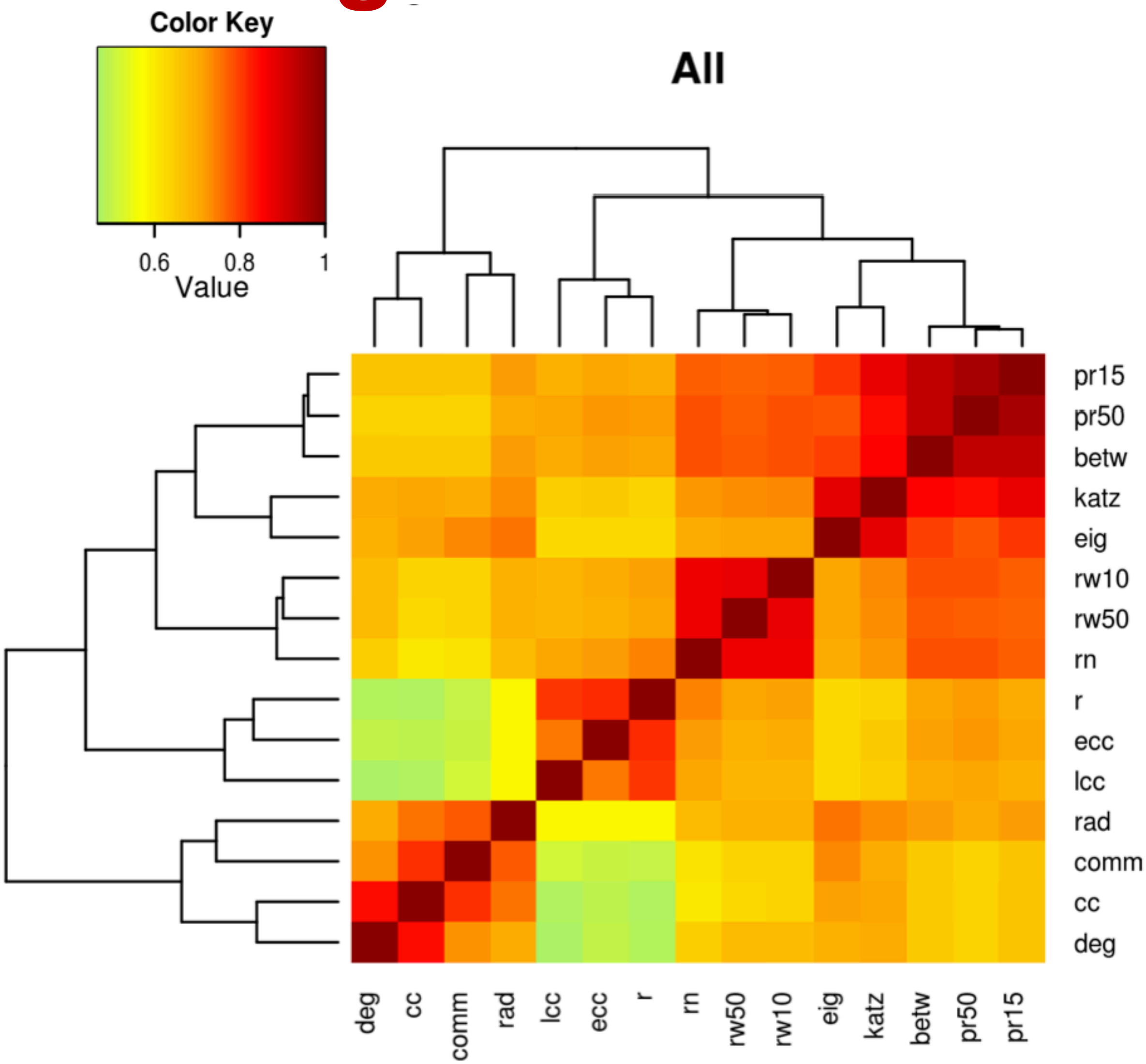
  social, bio, infra, info

- 3 analysis approaches

  I) rankings, II) node types, III) graph disruption

- Analysis results

Average-over-all heatmap



Pairwise similarity
Avg (Std)

BI-MATCH
0.88 (0.09)

SVM-Sep
0.69 (0.22)

Weight-Tau
0.36 (0.30)

- Weight-Tau in [-1,1]
- Others in [0,1]

BI-MATCH similarities (SVM and Weight-Tau are similar.)

Correlation Analysis

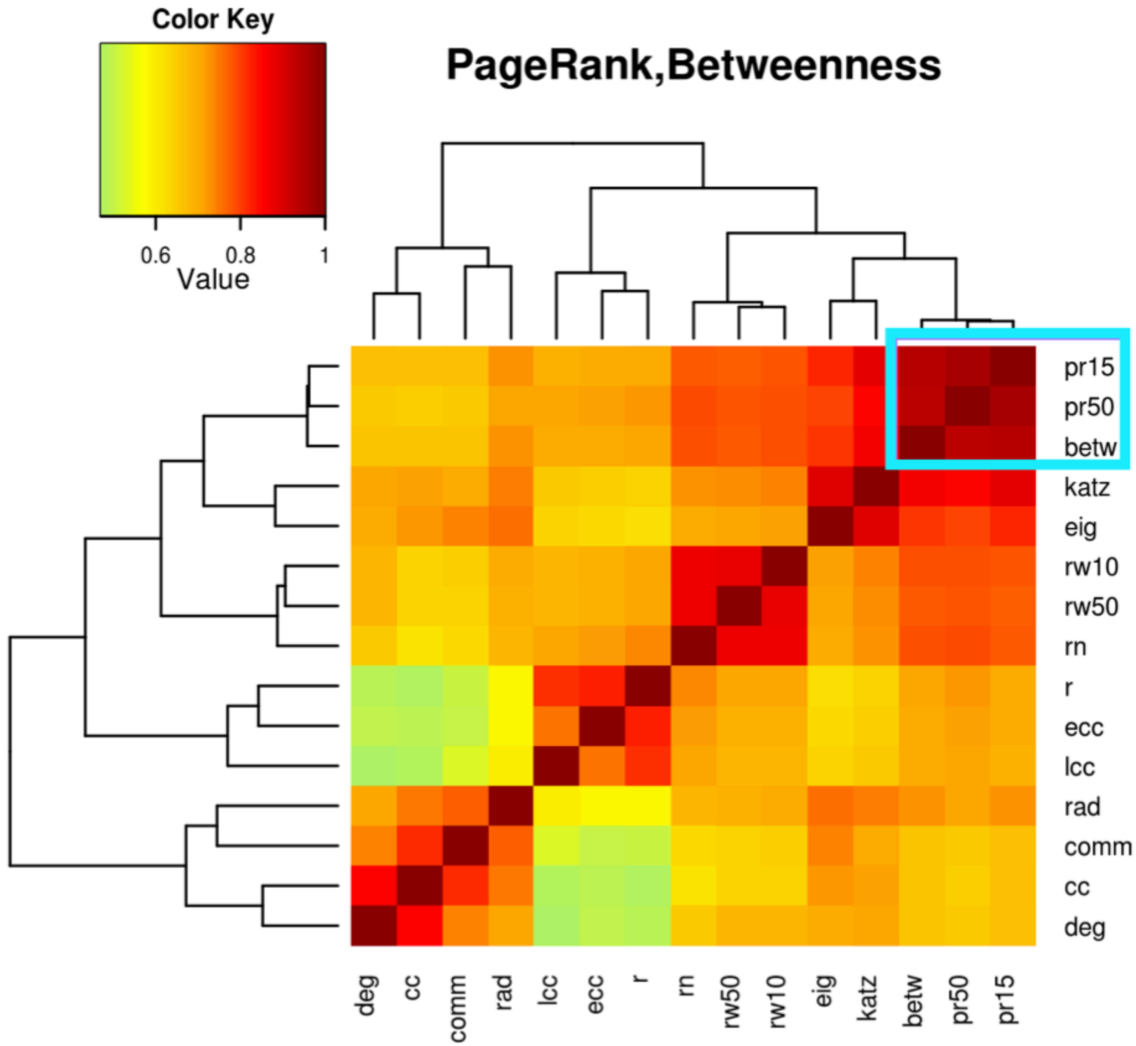
Stony Brook University
Computer Science

Clusters (Weight-Tau)	# Graphs
1. {PageRank15, PageRank50, Betweenness}	(67/68)
2. {Katz, Eigen-vector}	(56/68)
3. {Closeness, Communicability}	(40/68)
4. {Degree, Radius}	(39/68)

Clusters (SVM-Sep)	# Graphs
1. {PageRank15, PageRank50, Betweenness}	(64/68)
2. {Katz, Eigen-vector}	(54/68)
3. {Closeness, Degree}	(35/68)

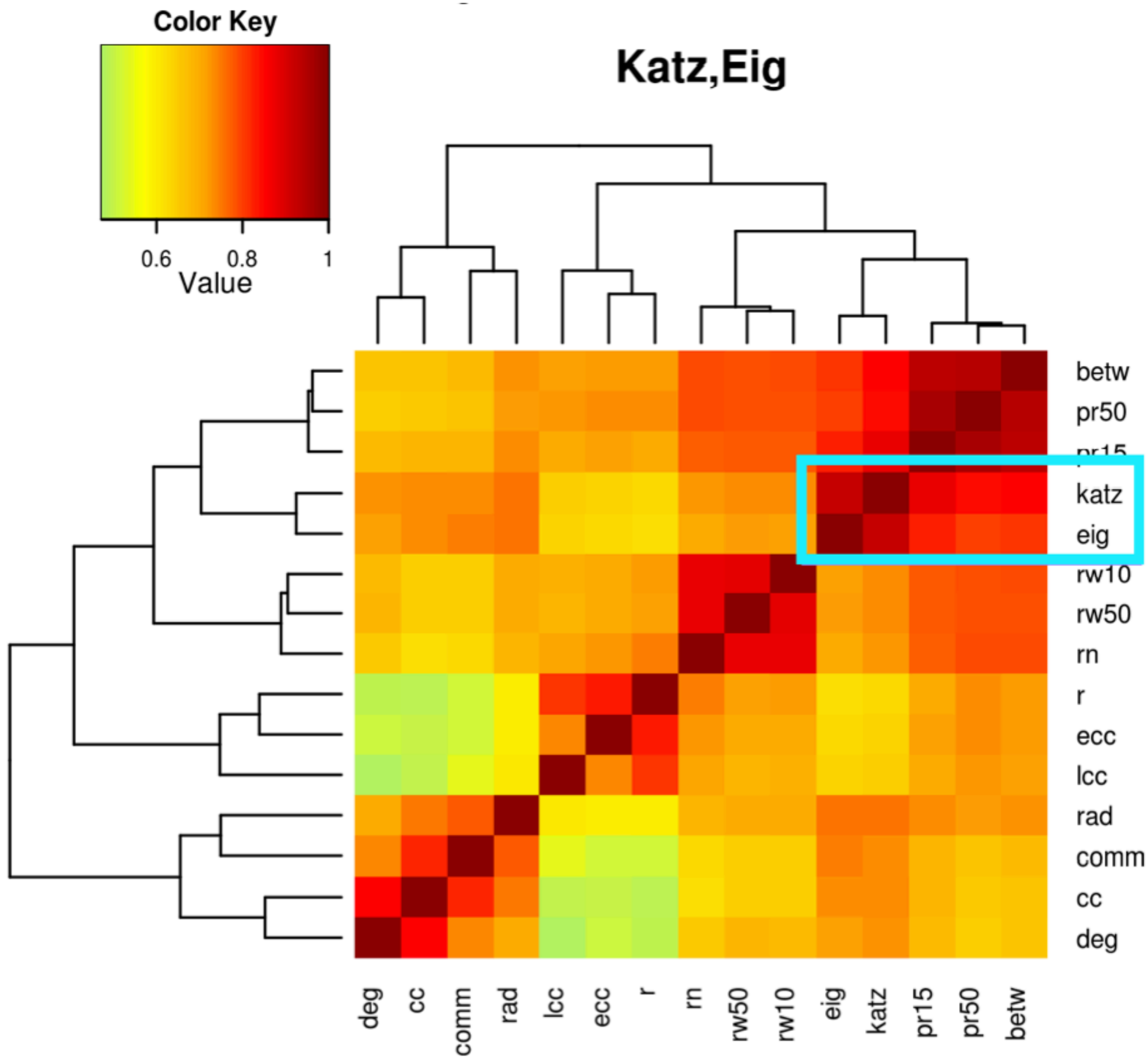
Clusters (Bi-Match)	# Graphs
1. {PageRank15, PageRank50, Betweenness}	(62/68)
2. {Katz, Eigen-vector}	(52/68)
3. {Closeness, Degree}	(44/68)

Group#1: Pagerank,Betweenness

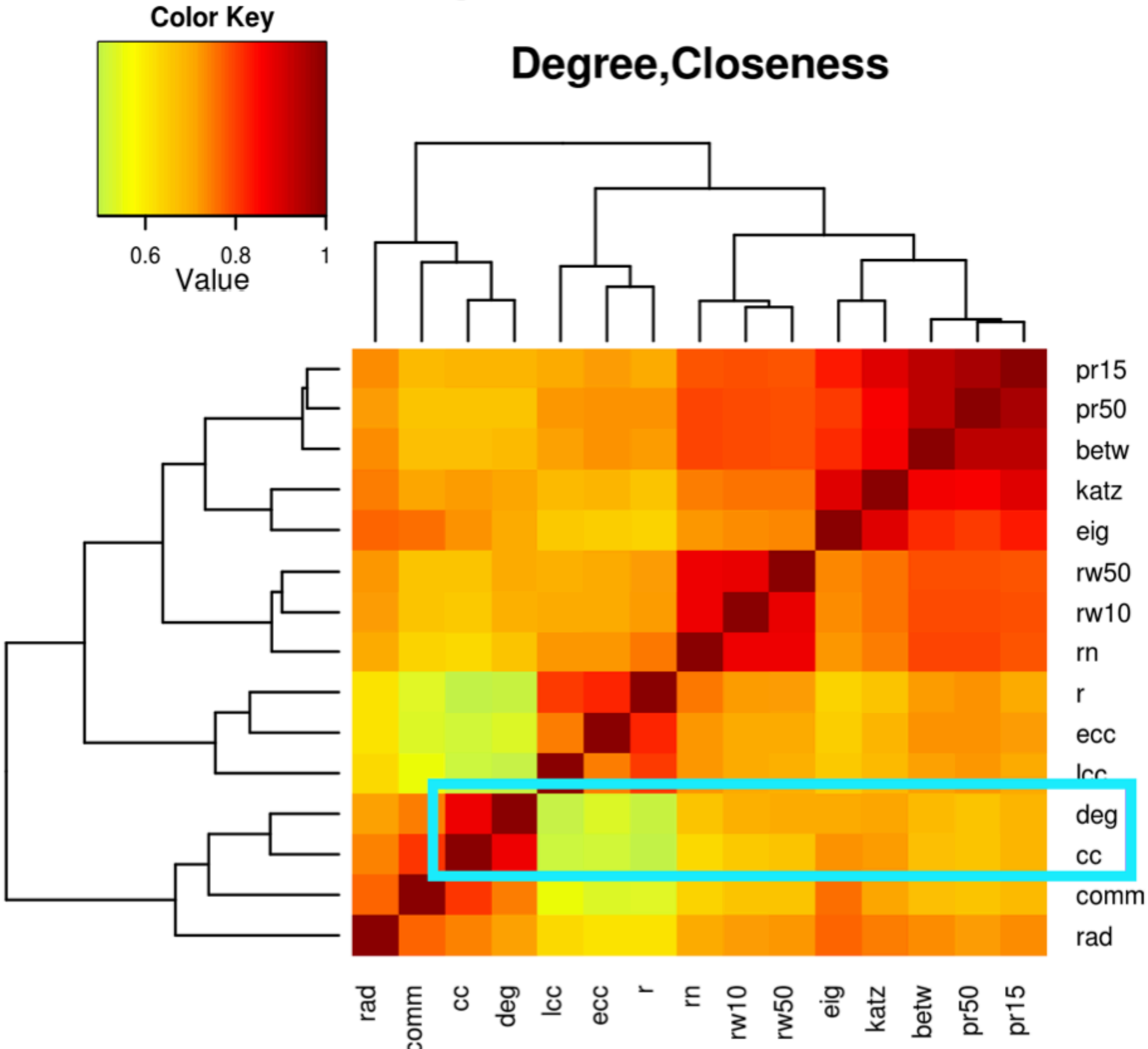


Pagerank: a cheap proxy to Betweenness

Group#2: Katz, Eigenvector

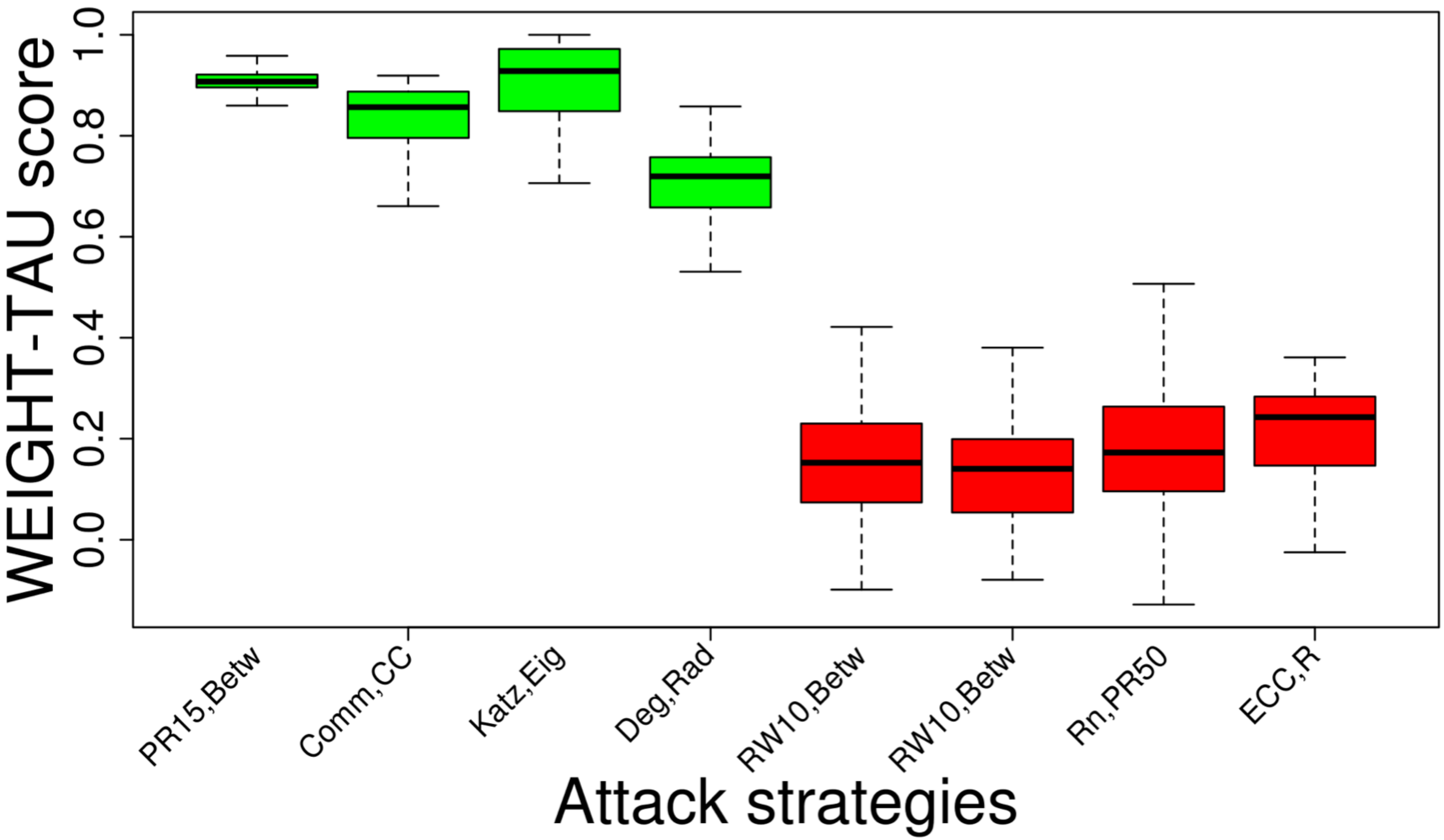


Group#3: Degree, Closeness



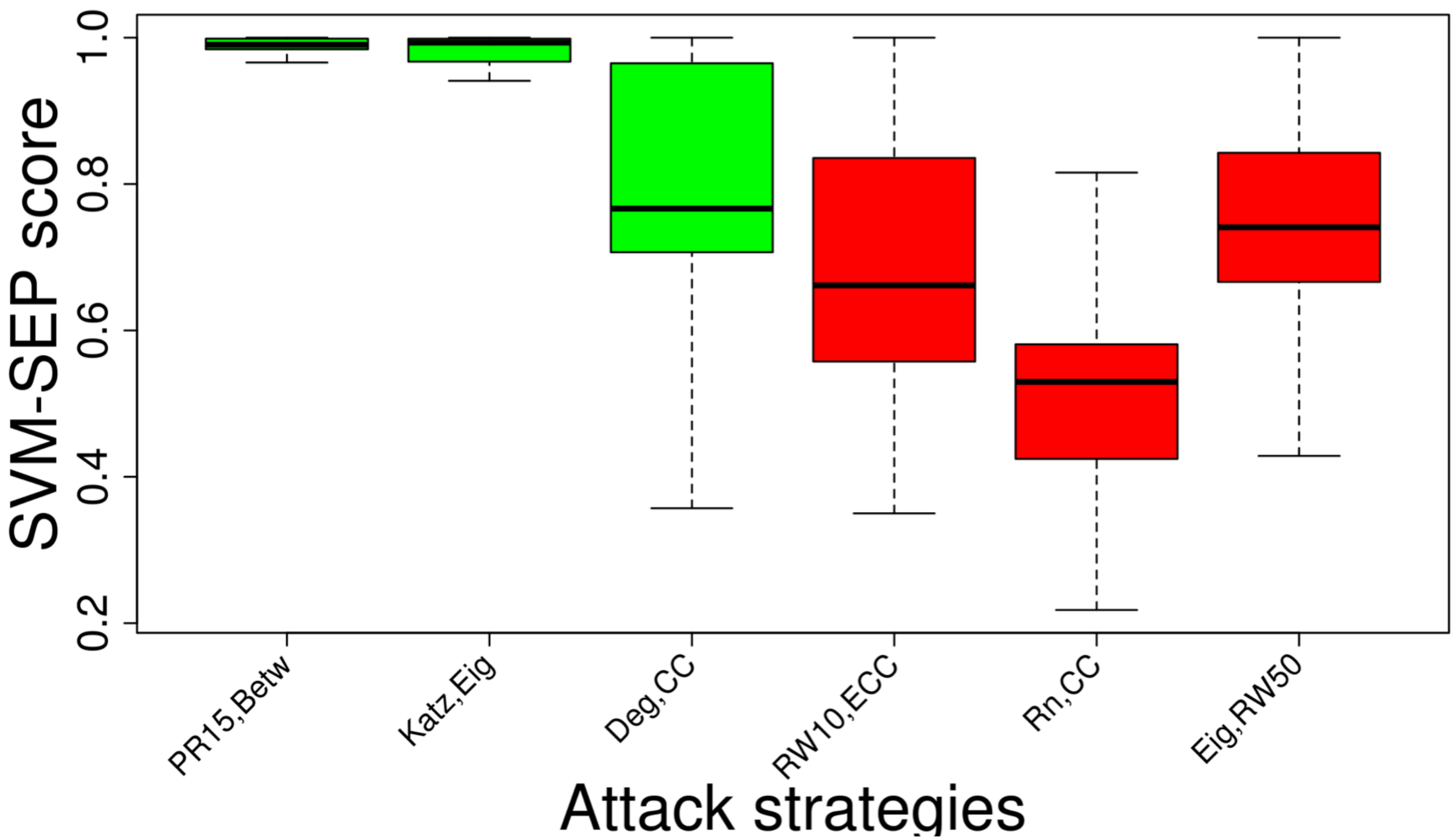
Degree: a cheap proxy to Closeness

Significance of correlation

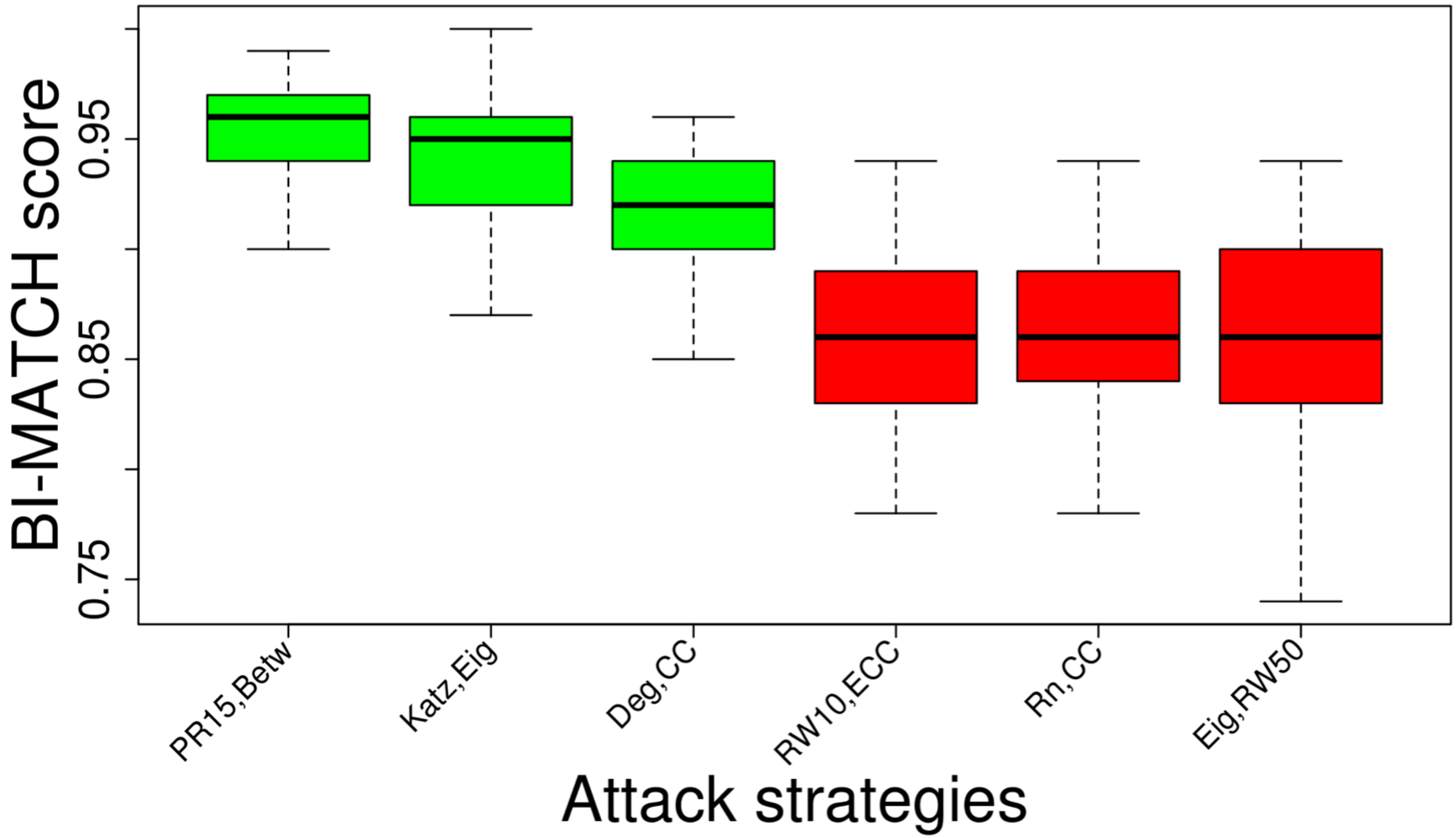


Box-plots: distribution of correlation across graphs in G

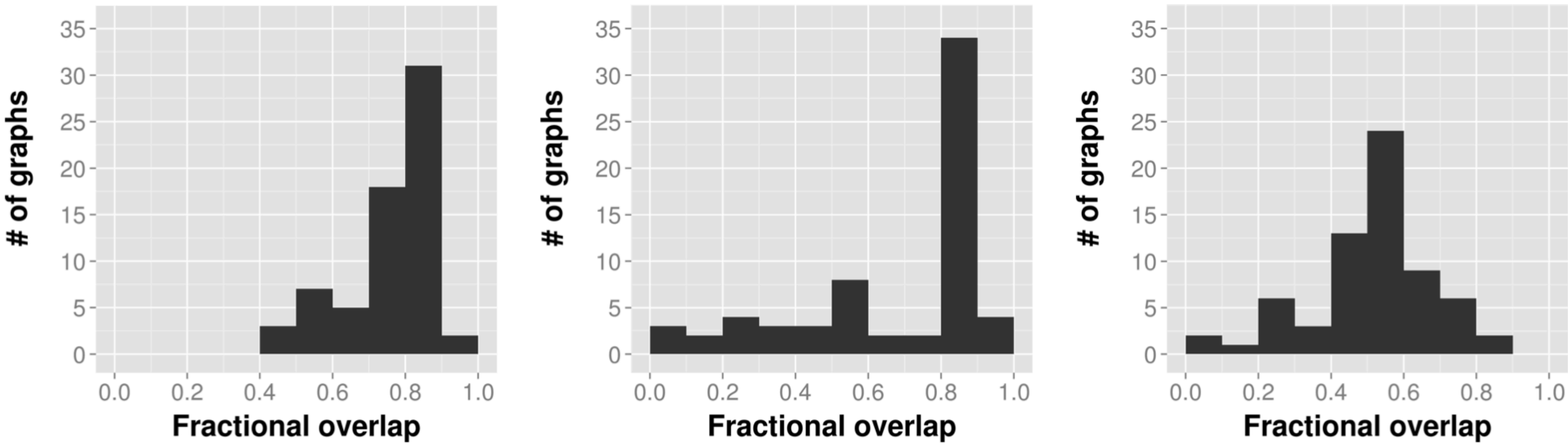
Significance of correlation



Significance of correlation



Overlap ratio of top-k nodes



(a) pr15 and betw (b) katz and eig (c) deg and cc

k is set to number of 1% of nodes in each graph

Consensus analysis

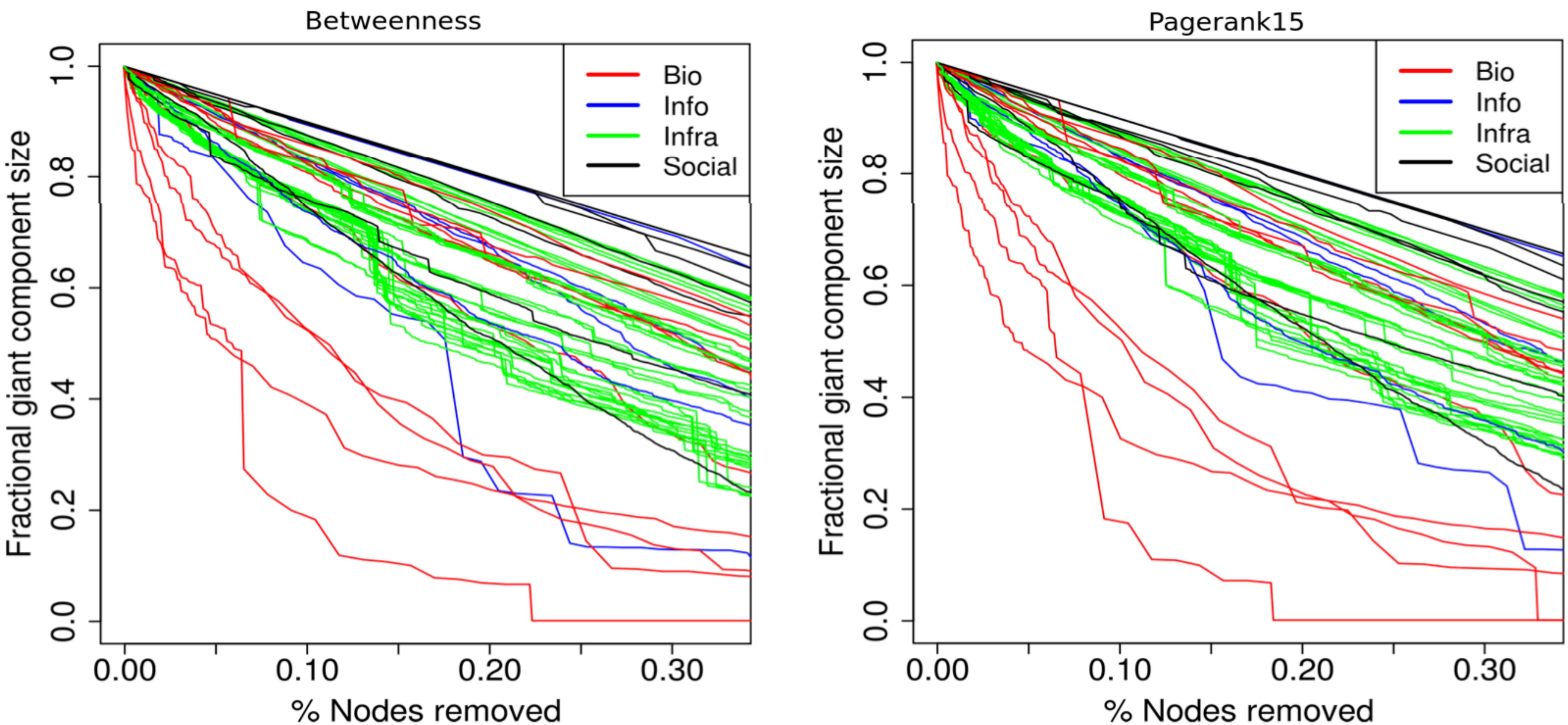
- Compute Kemeny-Young consensus on RANK-C ranking of nodes
- Sort strategies by closeness to consensus

Top 5 node-based strategies closest to the Kemeny-Young consensus across 10 example graphs.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
katz	katz	pr15	katz	betw	pr15	pr15	katz	katz	pr15
eig	pr15	katz	pr15	katz	katz	katz	pr15	eig	katz
pr15	pr50	pr50	pr50	pr15	pr50	comm	eig	pr15	pr50
betw	eig	betw	cc	pr50	betw	deg	pr50	pr50	eig
ecc	deg	eig	comm	ecc	eig	eig	betw	deg	betw

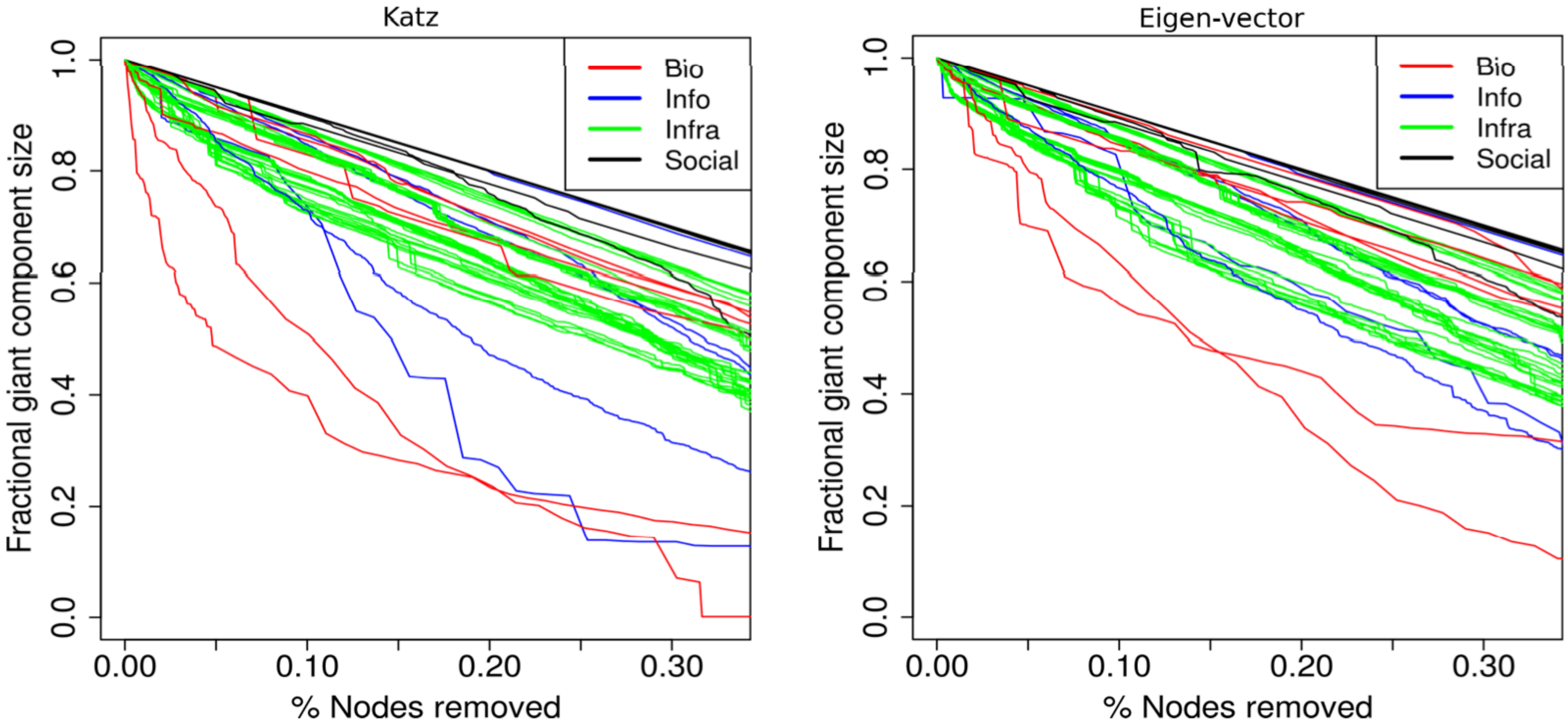
Katz or pr15 : cheap proxy to consensus

Disruption dynamics



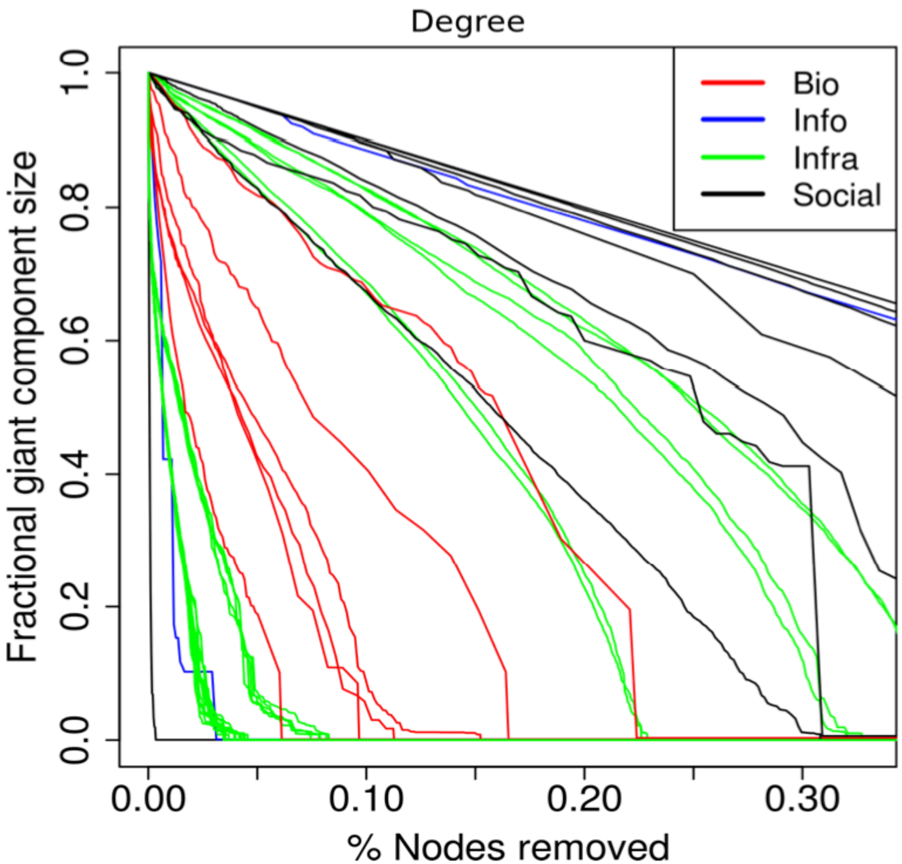
(a) Betwenness (betw) (b) PageRank15 (pr15)

Disruption dynamics

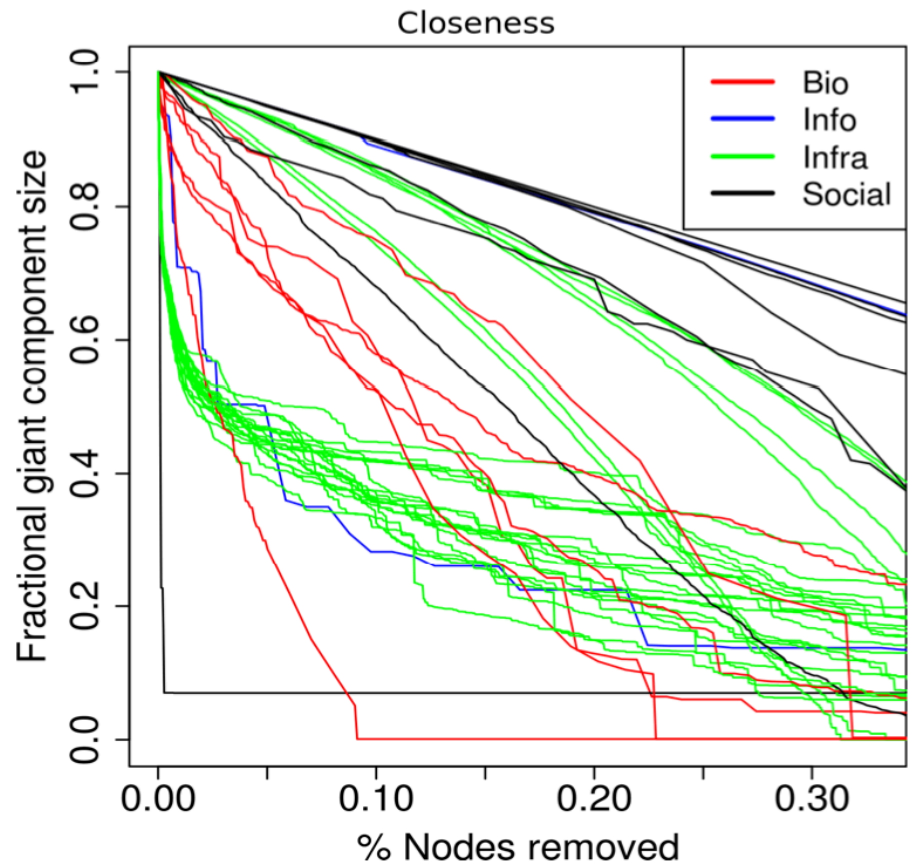


(c) Katz index (katz) (d) Eigen-vector (eig)

Disruption dynamics

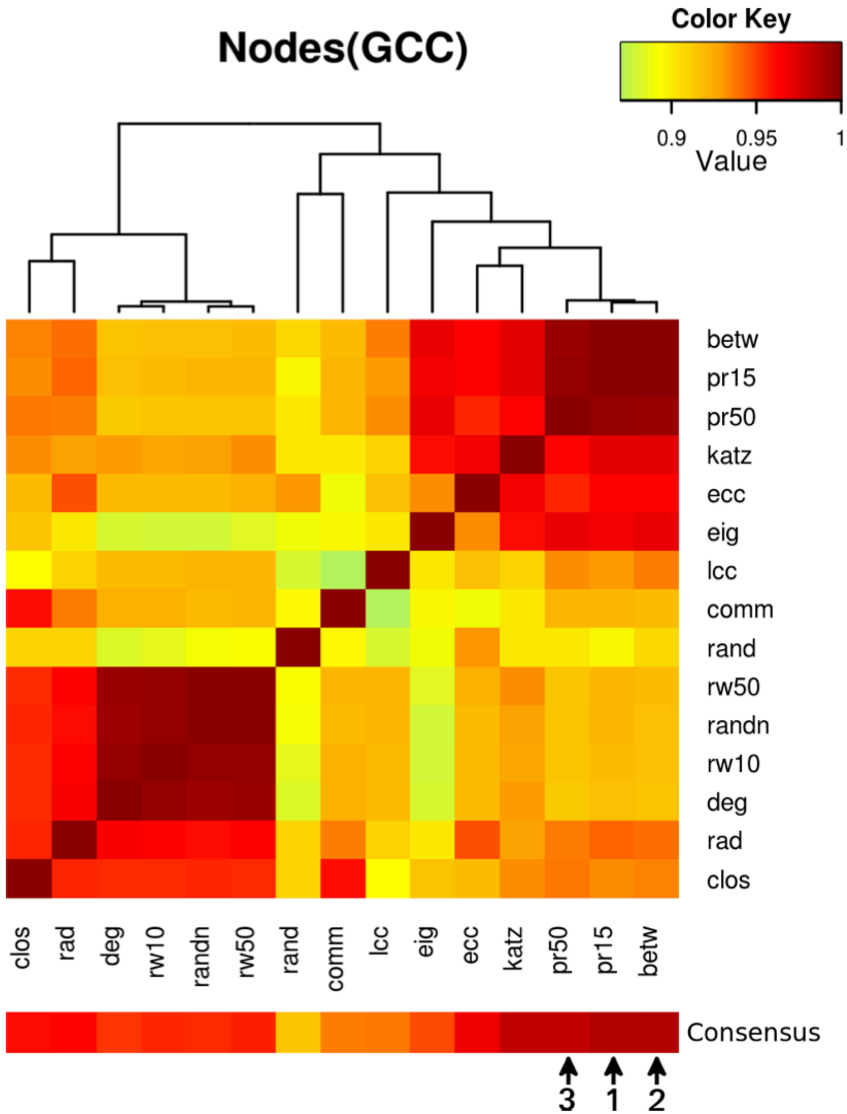


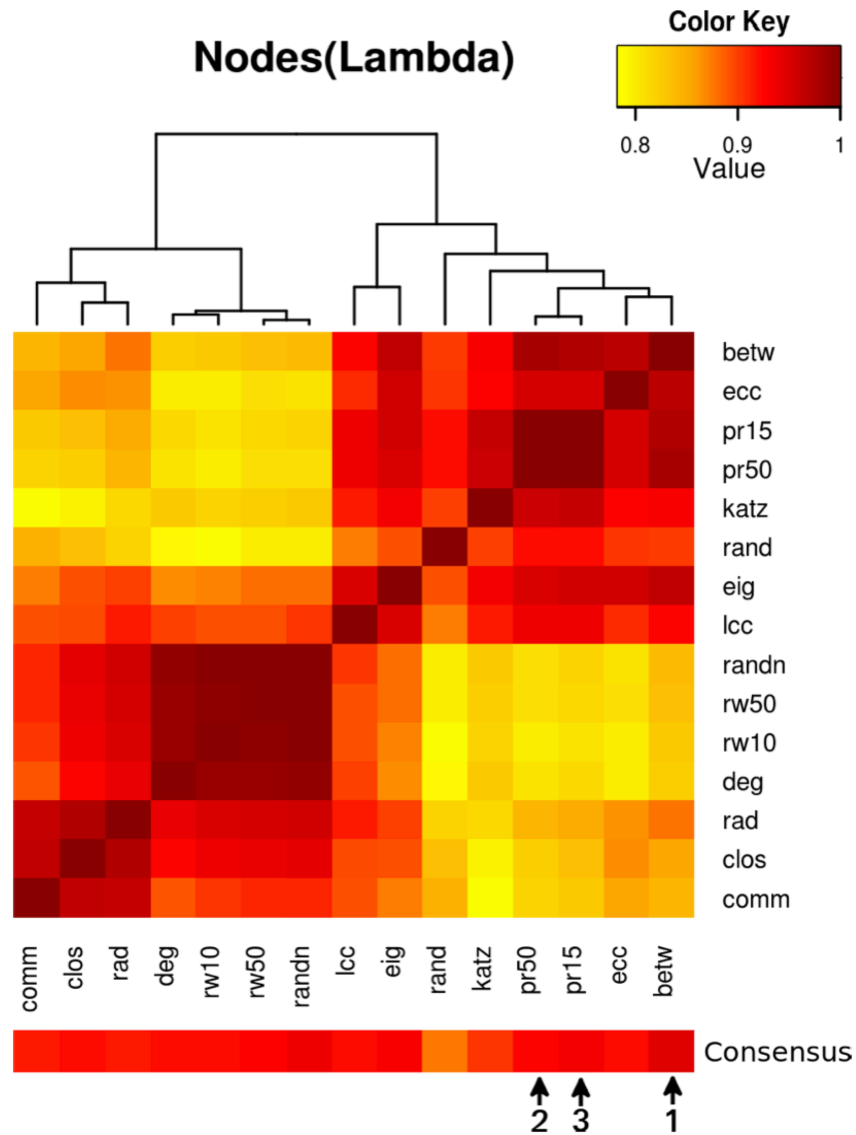
(e) Degree (deg)



(f) Closeness (cc)

Similarity by Response-C





Conclusion

Stony Brook University Computer Science

Summary:

- Studied of 15 measures, 68 graphs (4 domains)
- Employed 3 analysis approaches

Findings:

- High correlation across measures
- Significant groups of strongly correlated strategies (i.e., measures)

Implications:

- Cheap alternatives/approximation
- Proxy to consensus

Thank you!

leman@cs.stonybrook.edu

http://www.cs.stonybrook.edu/~datalab/