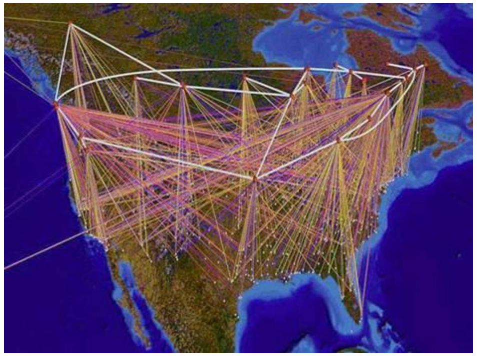
Comparative Assessment of Centrality Indices and Implications on the Vulnerability of ISP Networks

G. Nomikos[1], P. Pantazopoulos[1], M. Karaliopoulos[2], I. Stavrakakis[1]

[1] Department of Informatics and Telecommunications
National & Kapodistrian University of Athens
{gnomikos, ppantaz, ioannis}@di.uoa.gr

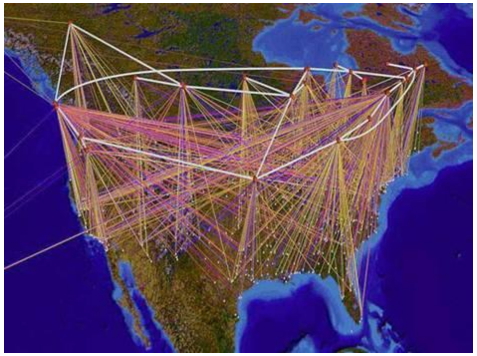
[2] Centre for Research and Technology-Hellas
Information Technologies Institute, Volos, Greece
mkaraliopoulos@iti.gr

Social Network Analysis: 'new' trend -'old' ideas



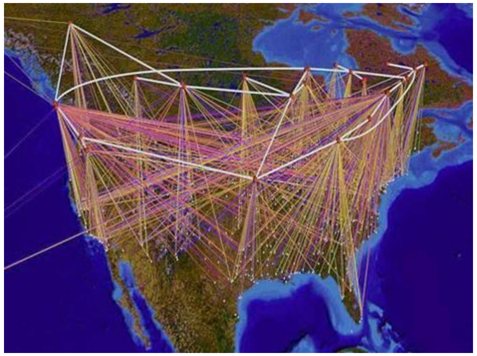
Social Network Analysis: 'new' trend -'old' ideas

- **SNA**: Interdisciplinary framework to process social info



Social Network Analysis: 'new' trend -'old' ideas

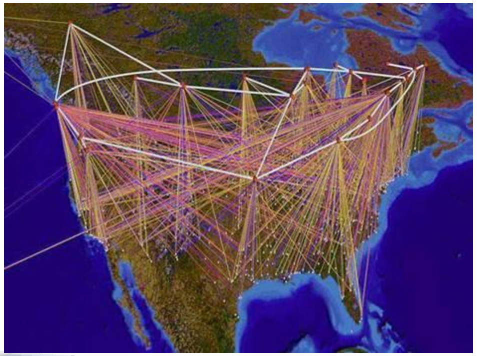
- **SNA**: Interdisciplinary framework to process social info



- Sociologists studied professional networks:

Social Network Analysis: 'new' trend -'old' ideas

- **SNA**: Interdisciplinary framework to process social info

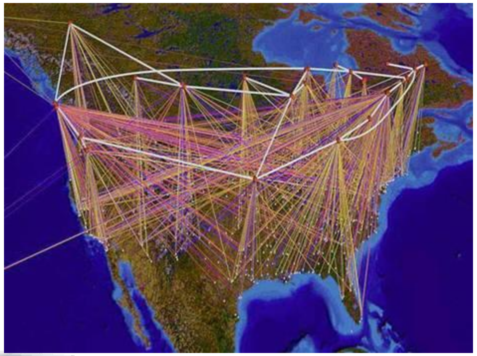


- Sociologists studied professional networks:
- Bavelas (1950) invented ***the notion of centrality***

to quantify the importance of an “actor” with respect to its social interconnections

Social Network Analysis: 'new' trend -'old' ideas

- **SNA**: Interdisciplinary framework to process social info



- Sociologists studied professional networks:
- Bavelas (1950) invented ***the notion of centrality***

to quantify the importance of an “actor” with respect to its social interconnections

- A multitude of centrality ***indices*** were proposed until late 70s
- Interest revived in late 90s mainly by the work of physicists
- Lately, centrality insights have been used for more efficient network protocol design

Studying the multiple instances of centrality

- **Motivation**: given that most of the proposed centrality indices are heuristic
  - How do they compare in their assessments about the relative importance of Internet nodes
  - Which one(s) may be “the right ones” for more reliable predictions of network vulnerability

- **Objectives**:
  - Study and classify the variety centrality indices proposed over the last fifty years
  - Assess the consistency of Internet node rankings induced by those indices
  - Compare indices with respect to their capacity to reveal Internet vulnerability to node attacks (*i.e.* removals)

A systematic study of the multiple centrality instances

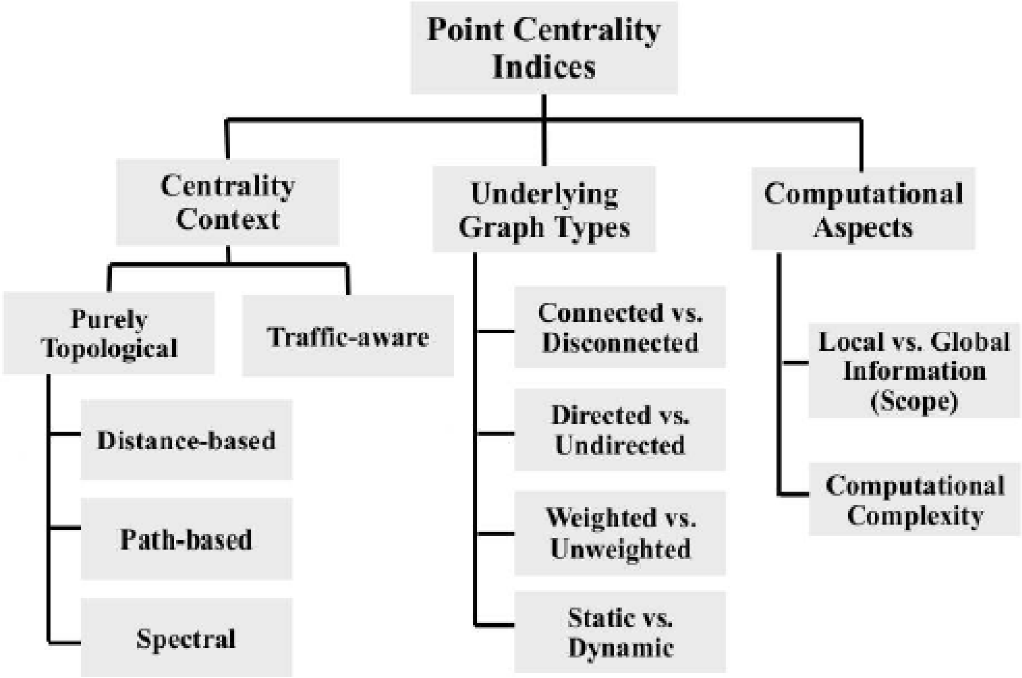
- Part I
- Thorough review and novel classification

- Part II
- Correlation study over a broad set of ISP network topologies

- Part III
- Impact of centrality-driven node removals on the connectivity and traffic-carrying capacity of ISP network topologies

A novel centrality classification scheme

- Characterize centrality indices along a number of attributes
- Similar classification for graph centrality indices



A novel centrality classification scheme

Seven popular centrality indices categorized under the proposed scheme

Betweenness (BC)	The extent to which i lies in shortest paths linking all network pairs.
Closeness (CC)	How fast i reaches all other network nodes in a connected graph.
Harmonic (HC)	How fast i reaches all other network nodes in a connected/disconnected graph.
Degree (DC)	Assign importance to i according to the number of its immediate neighbors.
Eccentricity (ECC)	i is important if its maximum distance to any node is close to the graph's radius.
Eigenvector (EC)	Assign importance to i if it has important neighbors.
PageRank (PG)	Assign importance to i proportionally to the importance of those pointing to i.

A novel centrality classification scheme

Seven popular centrality indices categorized under the proposed scheme

Betweenness (BC)	The extent to which i lies in shortest paths linking all network pairs.
Closeness (CC)	How fast i reaches all other network nodes in a connected graph.
Harmonic (HC)	How fast i reaches all other network nodes in a connected/disconnected graph.
Degree (DC)	Assign importance to i according to the number of its immediate neighbors.
Eccentricity (ECC)	i is important if its maximum distance to any node is close to the graph's radius.
Eigenvector (EC)	Assign importance to i if it has important neighbors.
PageRank (PG)	Assign importance to i proportionally to the importance of those pointing to i.

Centrality Index	**Context**				**Type of underlying graph**							**Computational aspects**		
	Topology aware			**Flow aware**	**Binary/ weighted**		**Directed/ Undirected**		**Dynamic**	**Connected/ Disconnected**		**Information (local/global)**		**Complexity**
	path	distance	spectral		B	W	D	U	D	C	D	L	G	
Betweenness (BC)	✓				✓	✓	✓	✓	✓	✓	✓		✓	$O(VE)$
Closeness (CC)		✓			✓	✓	✓	✓	✓	✓			✓	$O(V(logV)E)$
Degree (DC)	✓				✓	✓	✓	✓	✓	✓	✓	✓		$O(V^2)$
Eccentricity (ECC)		✓			✓	✓	✓	✓		✓			✓	$O(V(logV)E)$
Eigenvector (EC)			✓		✓	✓		✓		✓			✓	$O(V^3)$
Harmonic (HC)		✓			✓	✓	✓	✓		✓	✓		✓	$O(V(logV)E)$
PageRank (PG)			✓		✓		✓			✓	✓		✓	$\Omega(\frac{E^2}{ln(1/(1-d))})$

A quick look at correlation and network vulnerability studies

- Correlation between centrality indices
  - Linear correlations over synthetic graphs and a couple of real world topologies [1]
  - BC-DC correlation results over AS-level snapshots [2]

- Network vulnerability to centrality-driven attacks
  - Most of the studies concern synthetic graphs e.g. [3]
  - The impact of the attack is assessed only by connectivity measures [4]

[1] C.-Y. Lee, "Correlations among centrality measures in Complex networks." [Online].http://arxiv.org/abs/physics/0605220
[2] A. Vázquez et al., "Large-scale topological and dynamical properties of the internet," Phys. Rev. E, vol. 65, no. 6, 2002
[3] R. Albert, H. Jeong, and A.-L. Barabasi, "Error and attack tolerance of complex networks," Nature, vol. 406, no. 6794, 2000.
[4] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks," Phys. Rev. E, vol. 65, no. 5, 2002

A systematic study of the multiple centrality instances

- Part I
- Thorough review and novel classification

- Part II
- Correlation study over a broad set of ISP network topologies

- Part III
- Impact of centrality-driven node removals on the connectivity and traffic-carrying capacity of ISP network topologies

Internet router-level topologies

- `mrinfo` topologies (**76 -1300** #nodes, **14** snapshots)
  - Snapshots correspond to Tier-1 and Transit ISPs
  - Collected during 2004-2008 using a multicast discovering tool

- `Rocketfuel` topologies (**40 - 2000** #nodes, **9** snapshots)
  - Widely used in experimental studies
  - 800 vantage points as `traceroute` sources

- `Caida` topologies (**2000-82000** #nodes, **7** snapshots)
  - Collected during Oct. Nov 2011
  - `Traceroute` probes to randomly chosen destinations from 54 monitors worldwide

- `Topology Zoo` (**20 - 74** #nodes, **18** snapshots)
  - Capacitated snapshots reported directly by network operators of academic and research networks

Capturing associations between node rankings

- Spearman correlation coefficient

$$\rho_V(C_1, C_2) = 1 - \frac{6 \sum_{u \in V} (r_{C_1}(u) - r_{C_2}(u))^2}{|V|(|V|^2 - 1)}$$

  - rC_i :rank of each node when ordered according to centrality C_i
  - Values lie in [-1,1]

- Percentage overlap

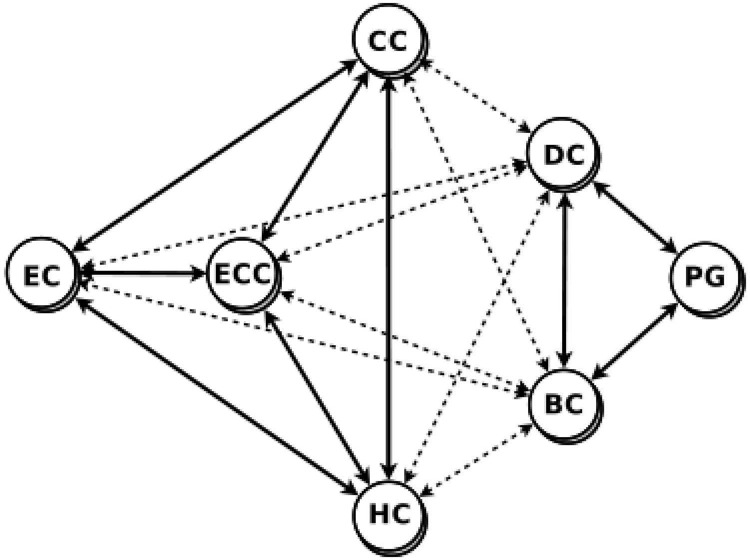
$$ov_V(C_1, C_2; k) = \frac{|\{v \in V : r_{C_1}(v) \leq k\} \bigcap \{v \in V : r_{C_2}(v) \leq k\}|}{k} \cdot 100\%$$

  - Overlap between the sets of the *k* most highly ranked nodes by the two centrality indices
  - Values lie in [0,100]

Rank- correlation strength results

- All centrality pairs are positively correlated

- Graph-based illustration of the rank correlation strength among the considered indices (similar trends across datasets)



- Solid lines:

Spearman coefficient ρ_v in [0.7-1]

- Dashed lines:

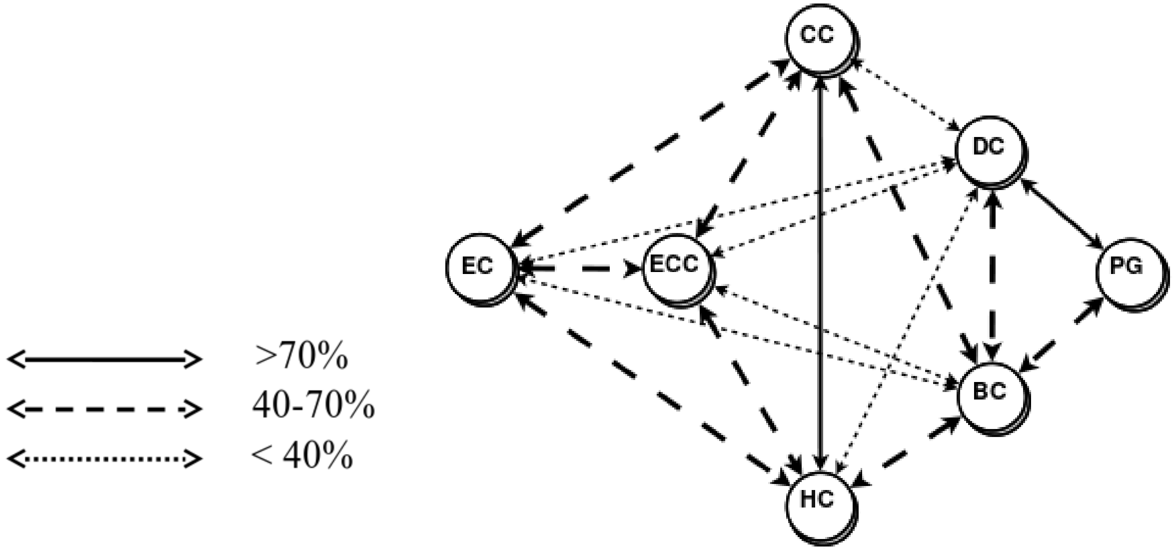
Spearman coefficients ρ_v in [0.3-0.7)

Some noteworthy relations

- Eccentricity vs. Closeness (*strong*)
  - It can be proved for trivial graphs (line, rectangular grid)
- Betweenness vs. Degree centrality (*strong*)
  - In agreement with earlier studies that report positive Pearson correlation over a wide range of networks
- PageRank vs. Degree centrality (*strong*)
  - PageRank has been shown to be statistically close to the degree centrality
- PageRank vs. Eigenvector (*weak*)
  - PG utilizes the damping factor *d* to determine the jump probability

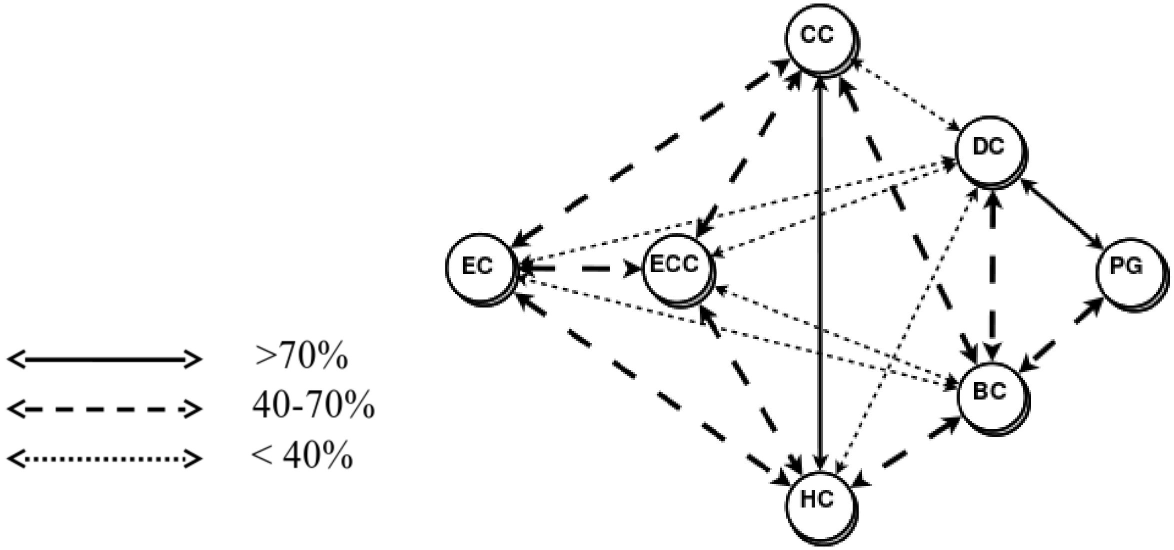
Top-5% overlap results

- Motivation: network protocols that seek to exploit a small set of the most central nodes
- Almost all centrality pairs found earlier to be strongly correlated appear more weakly associated in terms of overlap values
- Only two centrality pairs *i.e.*, PG-DC and HC-CC combine high overlap with strong rank correlation

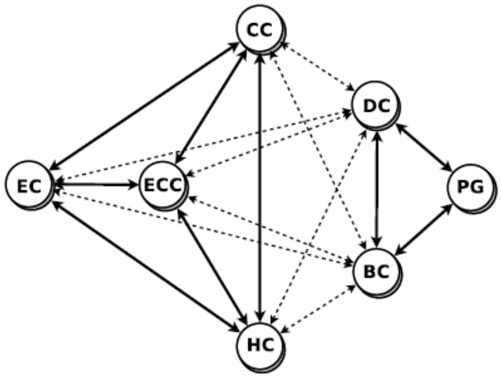


Top-5% overlap results

- Motivation: network protocols that seek to exploit a small set of the most central nodes
- Almost all centrality pairs found earlier to be strongly correlated appear more weakly associated in terms of overlap values
- Only two centrality pairs *i.e.*, PG-DC and HC-CC combine high overlap with strong rank correlation



rank-correlation graph



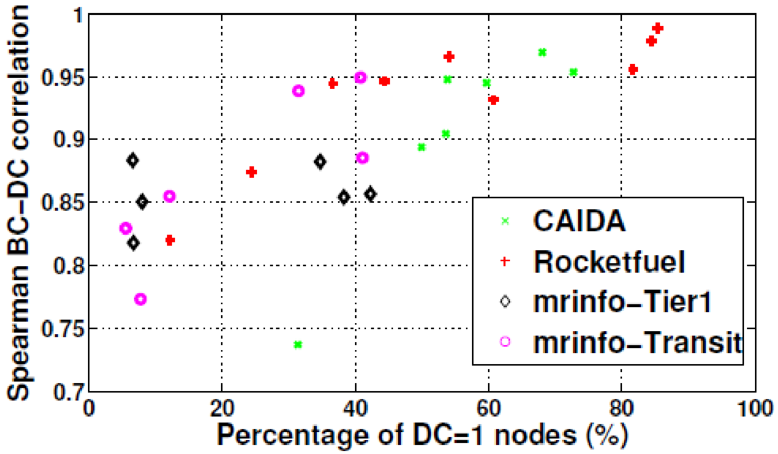
High correlation does not necessarily mean high top-k overlap

- BC-DC pair: nodes with the lowest DC value exhibit the lowest BC as well

- Nodes with DC=1 are expected to positively contribute to the BC-DC correlation

- The top DC and BC nodes do not coincide

- High correlation is mainly due to nodes of lowest rank

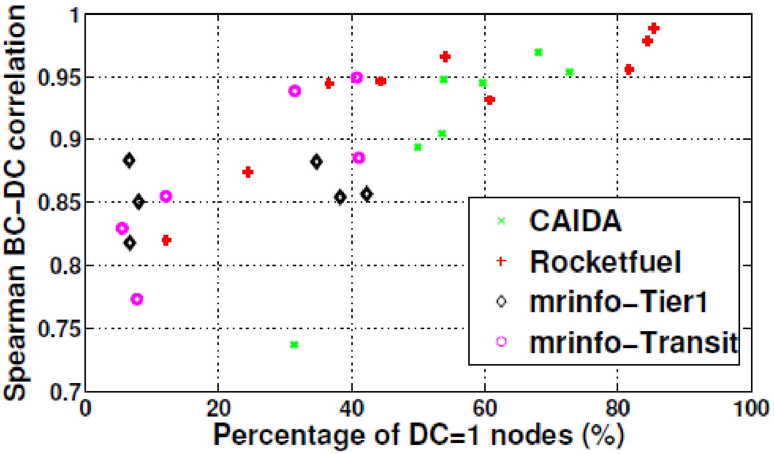


RANK CORRELATION STRENGTH VS. OVERLAP (%) BETWEEN BC AND DC

Dataset-ID	BC-DC Spearman Coefficient	Top-5% Overlap	Fraction of nodes having DC=1
CAIDA-1557	0.95	53%	54%
RocketFuel-1239	0.96	85%	82%
MrInfo, Tier1-1239	0.86	54%	43%
MrInfo, Transit-3292	0.94	40%	32%

High correlation does not necessarily mean high top-k overlap

- BC-DC pair: nodes with the lowest DC value exhibit the lowest BC as well
- Nodes with DC=1 are expected to positively contribute to the BC-DC correlation
- The top DC and BC nodes do not coincide
- High correlation is mainly due to nodes of lowest rank



RANK CORRELATION STRENGTH VS. OVERLAP (%) BETWEEN BC AND DC

Dataset-ID	BC-DC Spearman Coefficient	Top-5% Overlap	Fraction of nodes having DC=1
CAIDA-1557	0.95	53%	54%
RocketFuel-1239	0.96	85%	82%
MrInfo, Tier1-1239	0.86	54%	43%
MrInfo, Transit-3292	0.94	40%	32%

A systematic study of the multiple centrality instances

- Part I
- Thorough review and novel classification

- Part II
- Correlation study over a broad set of ISP network topologies

- Part III
- Impact of centrality-driven node removals on the connectivity and traffic-carrying capacity of ISP network topologies

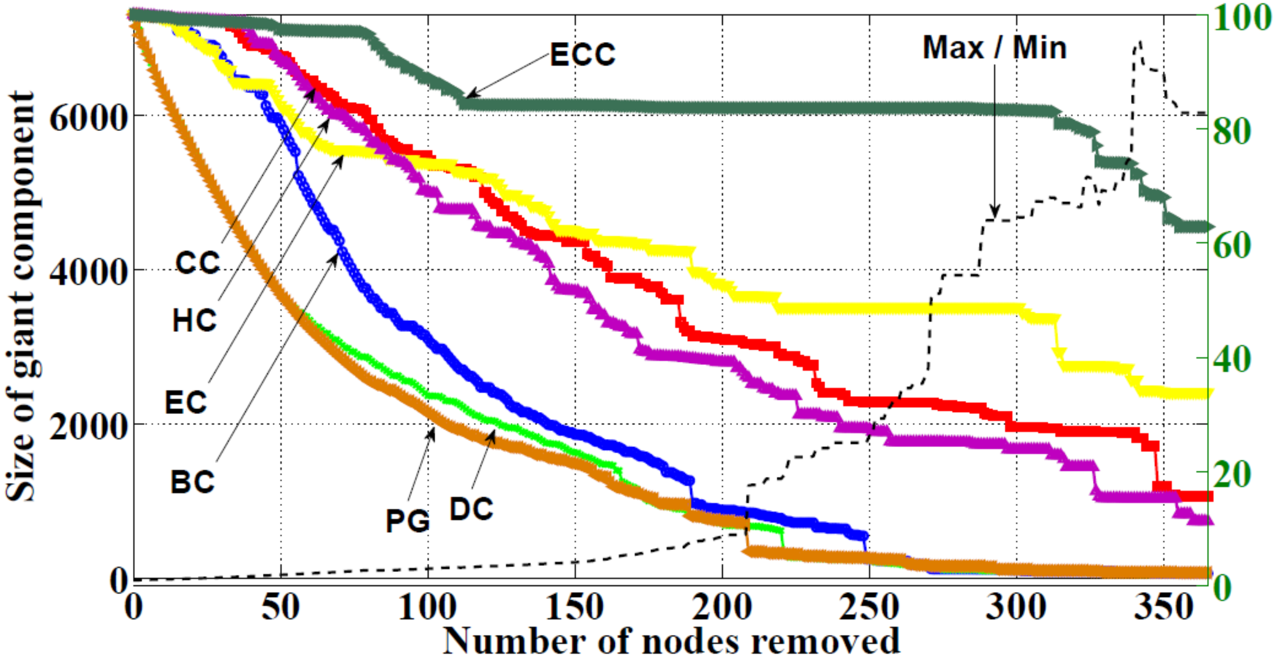
Centralities in the Internet vulnerability context

- We use the ranking that the seven considered metrics yield as criteria for Internet node attacks (*i.e.,* removals)

- We assess the impact of removing 5% of the most central network nodes in terms of
    - Network Connectivity
        - Giant Connected Component (GCC)
        - Number of connected components
        - Average shortest-path length

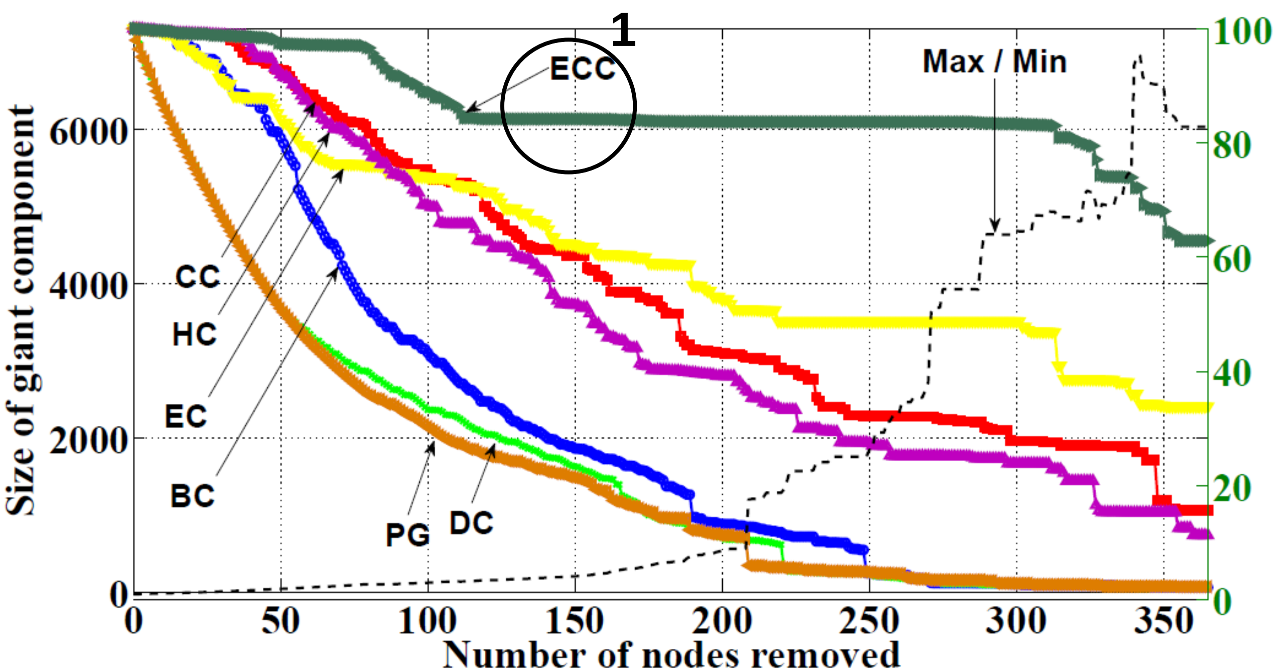
- Network "Throughput"
    - Traffic-serving capacity of the network

Connectivity results: Giant Connected Component



Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

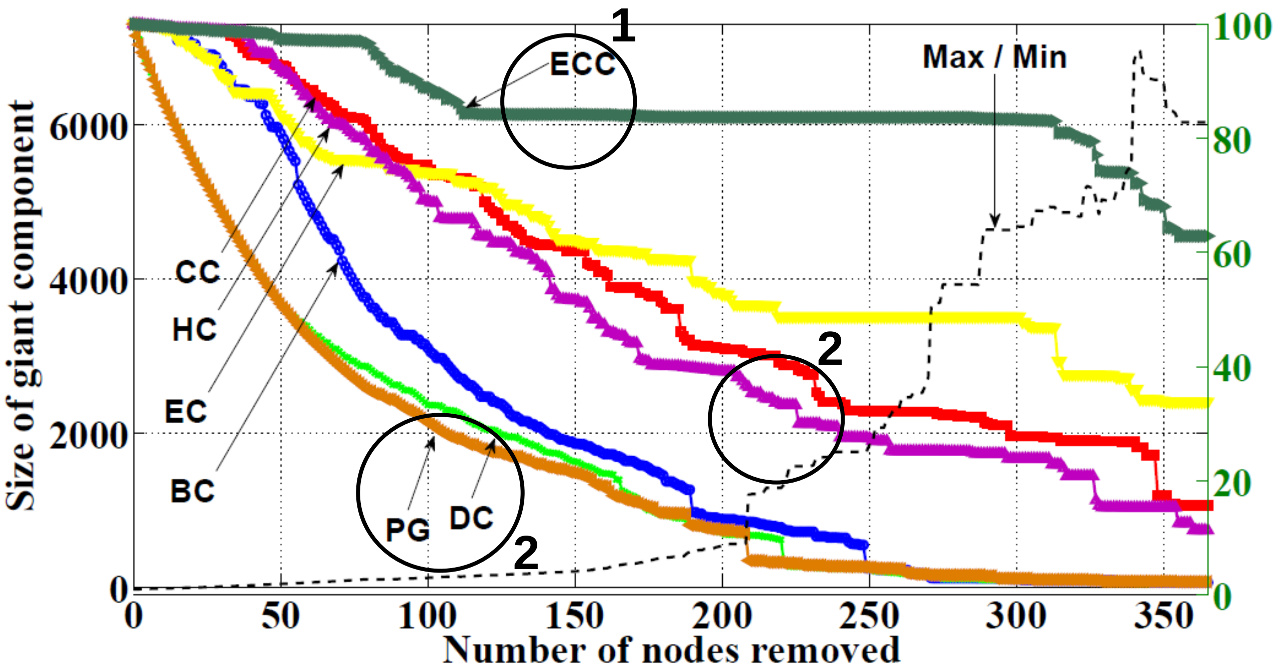
Connectivity results: Giant Connected Component



1. **1.** ECC, the index with the less dramatic impact

Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

Connectivity results: Giant Connected Component

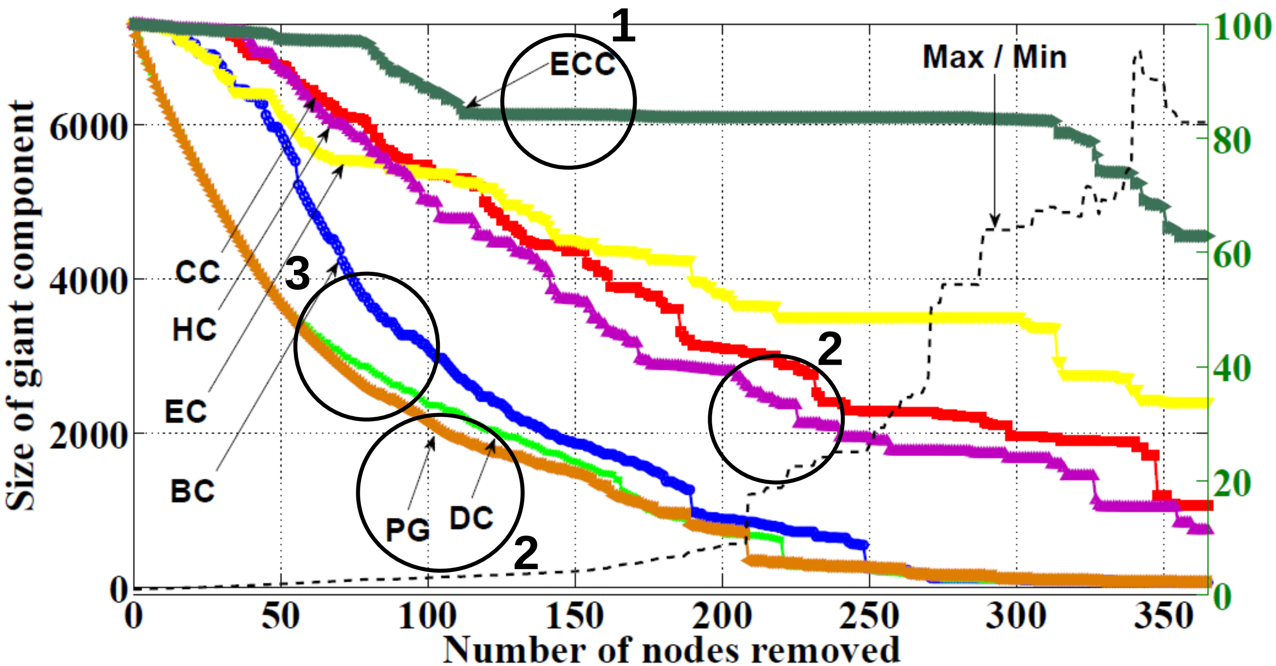


1. ECC, the index with the less dramatic impact

2. PG-DC, CC-HC in good agreement with the strong correlation earlier observed

Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

Connectivity results: Giant Connected Component



Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

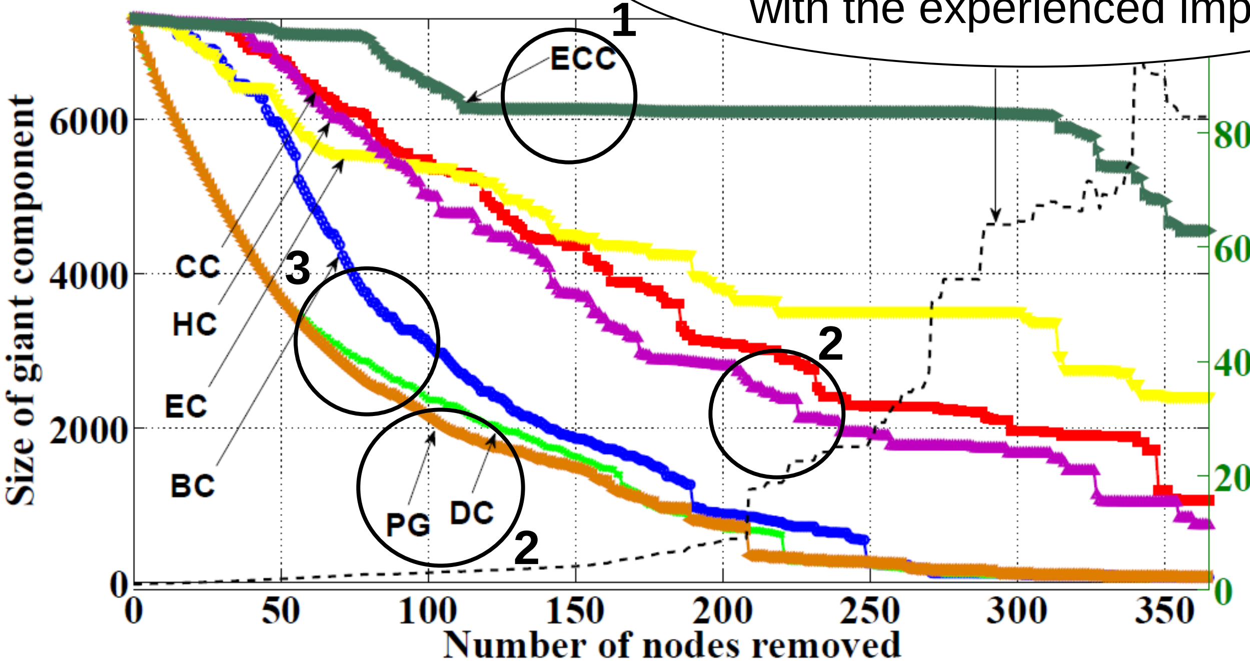
1. ECC, the index with the less dramatic impact

2. PG-DC, CC-HC in good agreement with the strong correlation earlier observed

3. BC-DC, dissimilar impact *w.r.t.* the rank-correlation and top-5% overlap

Connectivity results: Giant Connected Component

The full-rank correlation values are not always in line with the experienced impact



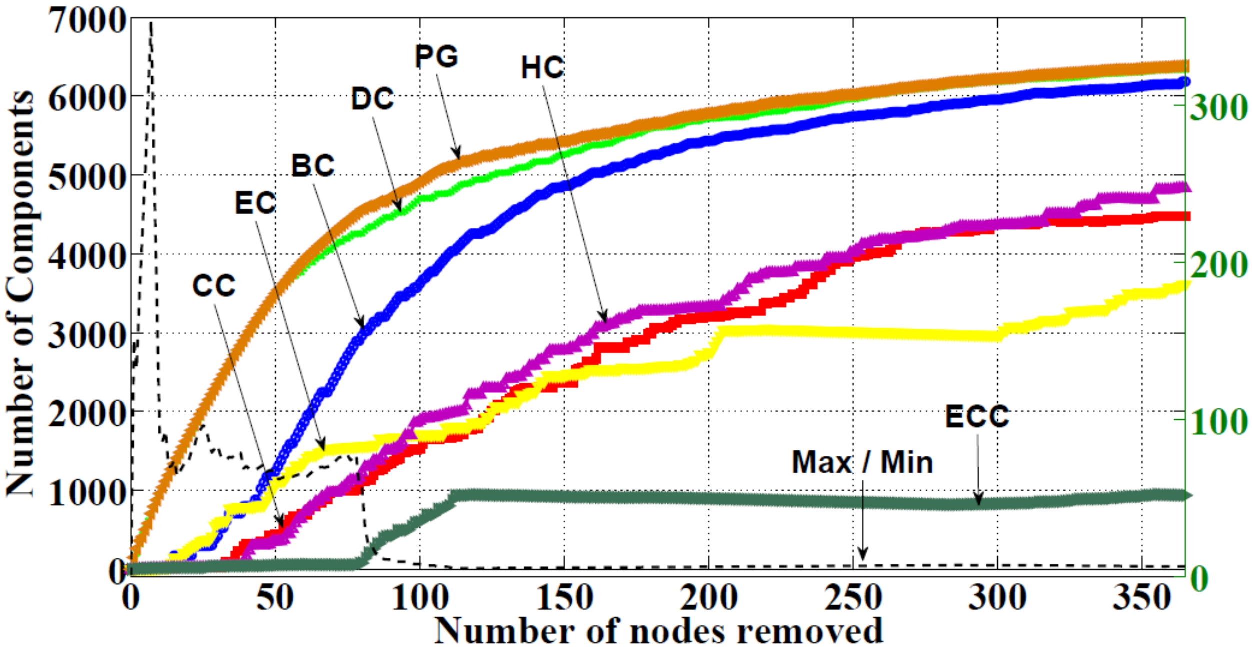
1. ECC, the index with the less dramatic impact

2. PG-DC, CC-HC in good agreement with the strong correlation earlier observed

3. BC-DC, dissimilar impact *w.r.t.* the rank-correlation and top-5% overlap

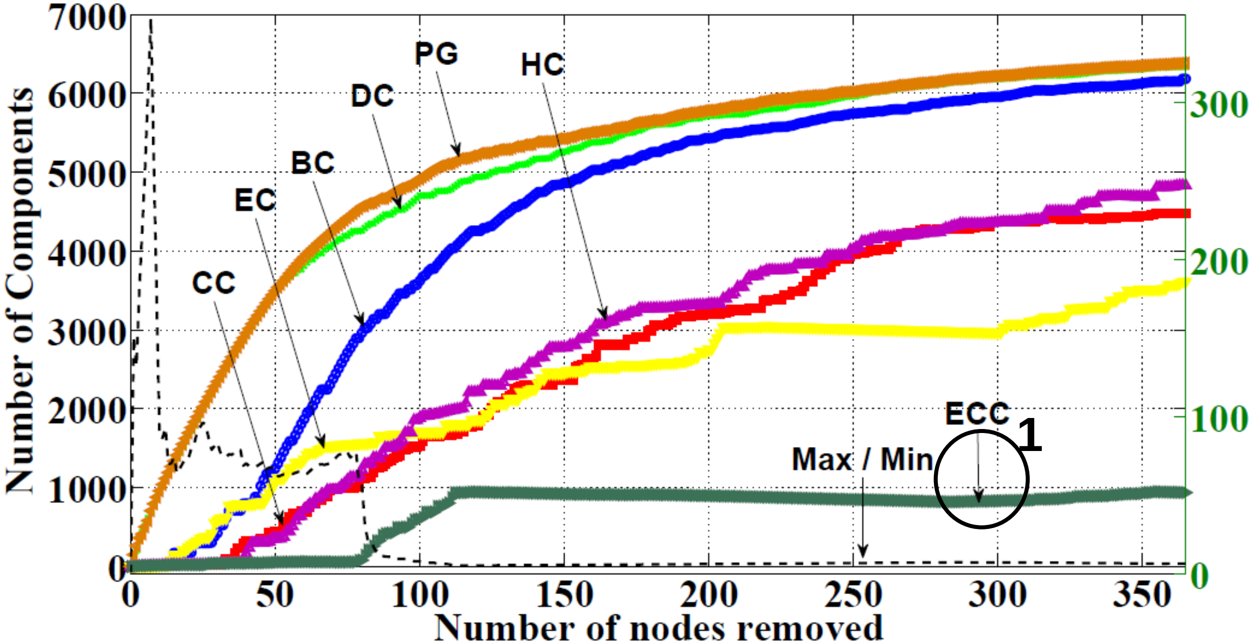
Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

Connectivity results: Number of Connected Components



Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

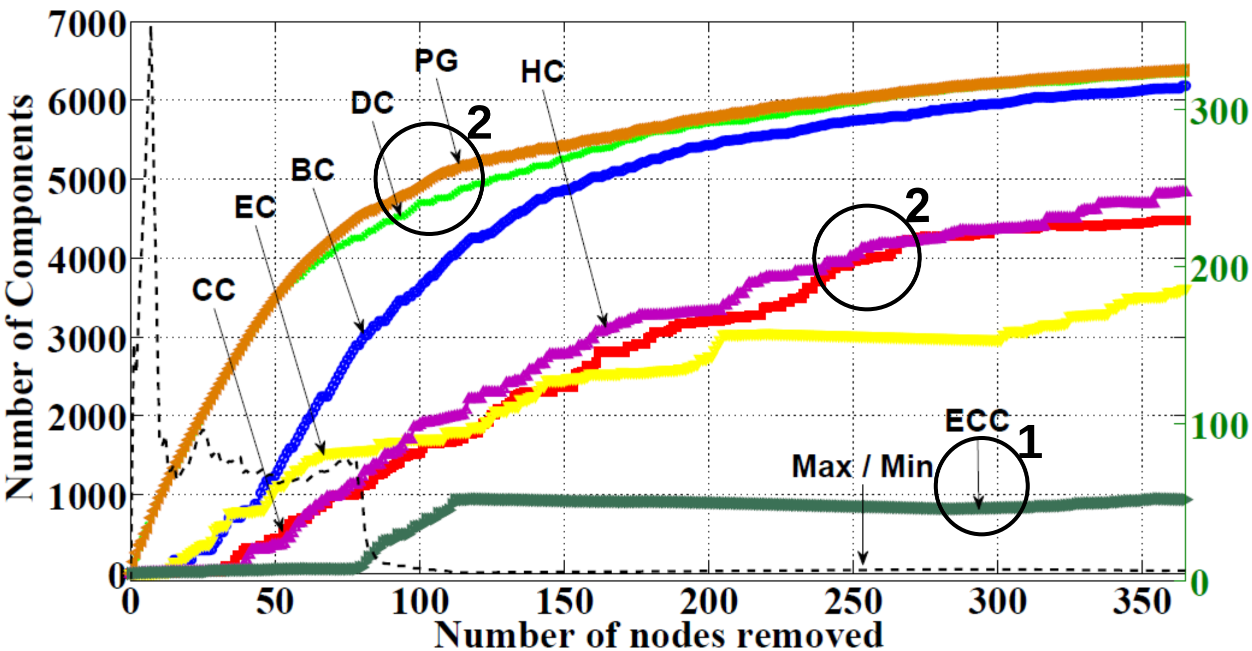
Connectivity results: Number of Connected Components



1. ECC, same behavior as previously observed

Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

Connectivity results: Number of Connected Components

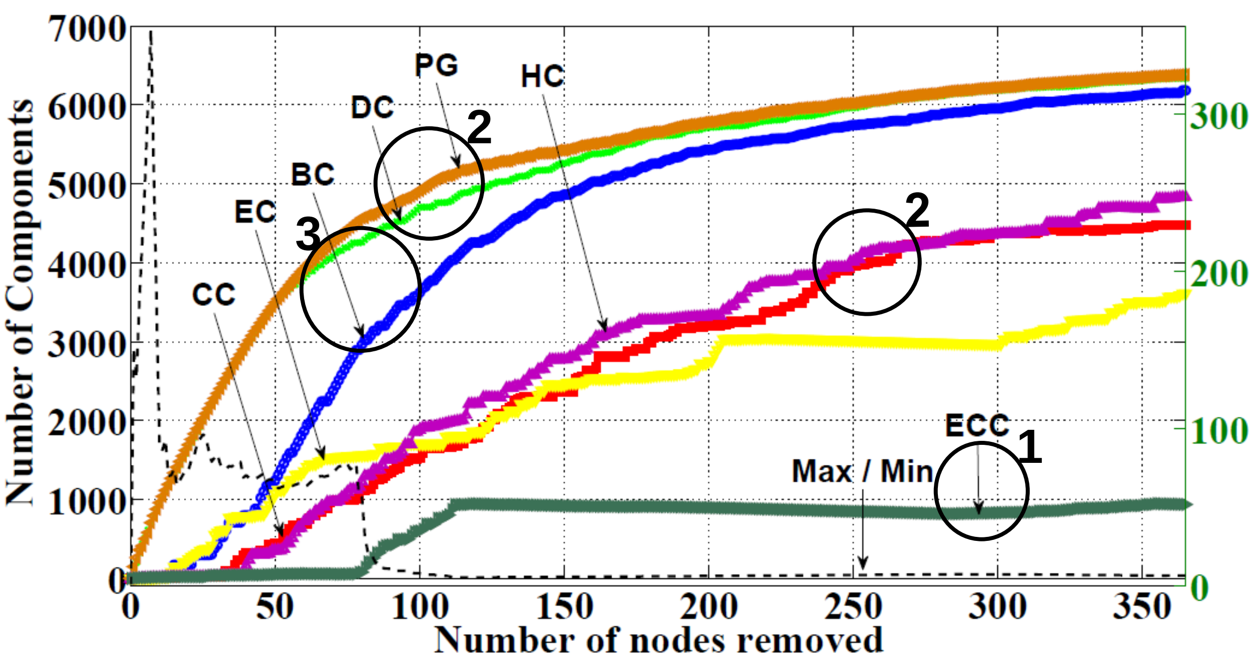


1. ECC, same behavior as previously observed

2. PG-DC (dominant in terms of partitioning), CC-HC in good agreement with the strong correlation earlier observed

Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

Connectivity results: Number of Connected Components



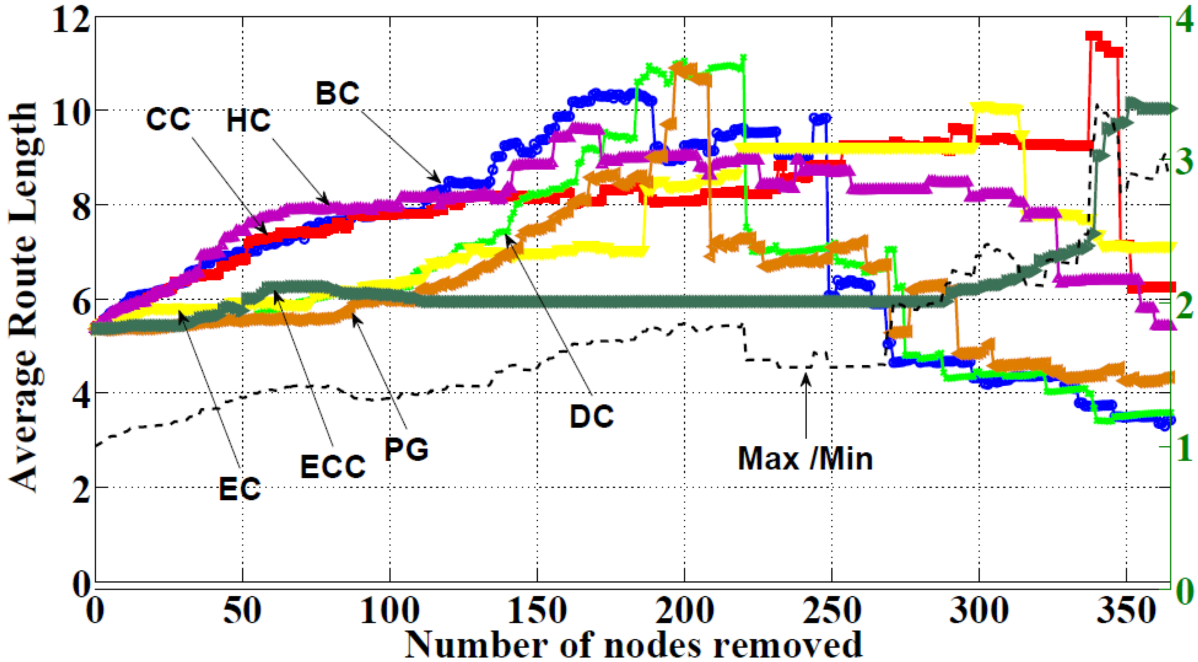
Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

1. ECC, same behavior as previously observed

2. PG-DC (dominant in terms of partitioning), CC-HC in good agreement with the strong correlation earlier observed

3. BC-DC, dissimilar impact *w.r.t.* the rank-correlation

Connectivity results: Average shortest-path length



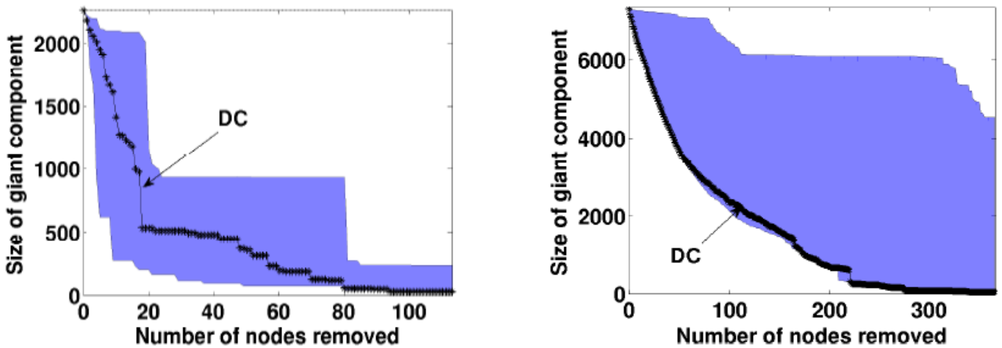
Average shortest-path length does not offer a clear view

A twofold behavior: Average path increases up to a point; further removals create single isolated nodes

Rocketfuel - AS1239, Size: 7303 – 365 removed nodes (5%)

Connectivity results: local vs. global indices

- The removals of the most central nodes affect differently the network connectivity
- Envelope plot to mark the best- and worst-case for the connectivity metric



a. Caida-AS786

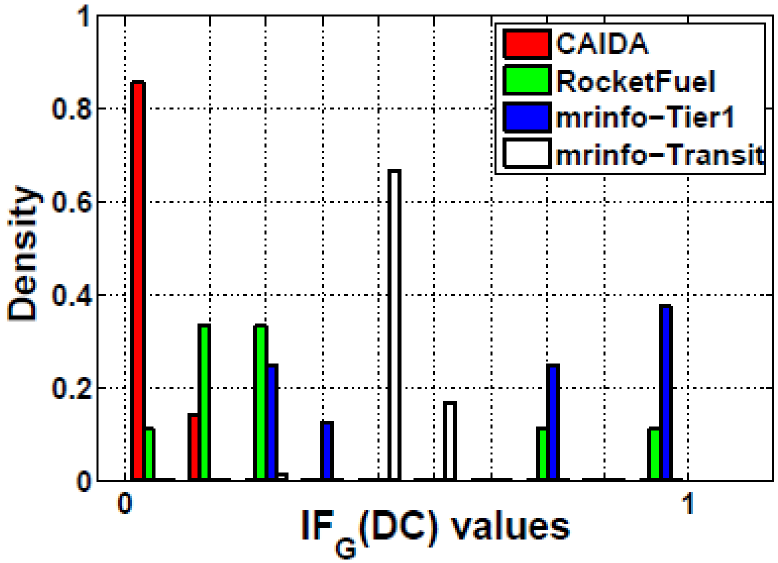
b. Rocketfuel-AS1239

- Where in this envelope the metric values corresponding to Degree Centrality, lie?
- Normalized distance metric over network G for centrality c and a set K of removed nodes

$$IF_G(c) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \frac{|m(k;c) - m_{wc}(k)|}{|m_{bc}(k) - m_{wc}(k)|}$$

Connectivity results: local vs. global indices

- $IF_G(k;c)$ takes values in [0,1]
- Plot the empirical probability mass of IF_G



- Over CAIDA, DC closely approximates the global metric with the most dramatic connectivity impact
- Over the other two datasets, DC cannot offer an effective approximation

Centrality-driven node removals and "Network throughput"

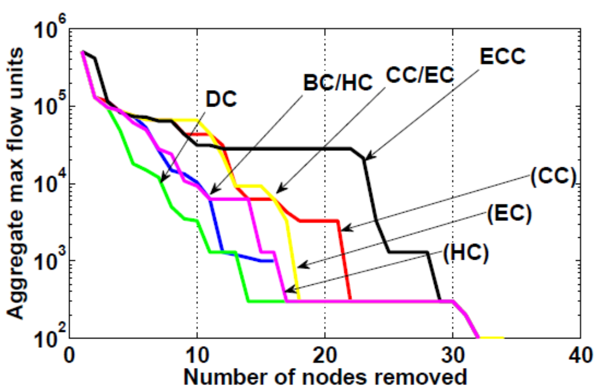
- Impact of centrality-driven node removals on the network traffic serving capacity

- One way to do it: Employ traffic matrices and then solve the Multicommodity Flow Problem
- Limitations:
  - Traffic matrices are rarely known a-priori,
  - Variations over time
  - MCF problem is NP-complete

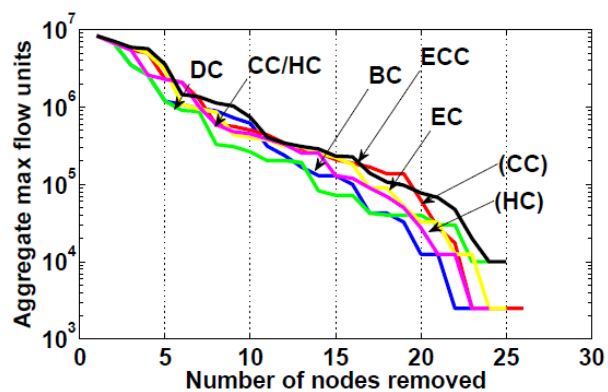
- A simpler approach: Compute the sum of maximum flows over all network node pairs
  - This sum is a very loose upper bound of traffic load that can be simultaneous served by the network

“Network throughput” results

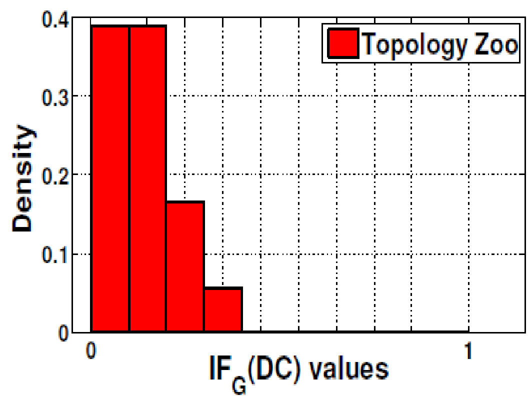
- Reduction in the aggregate max-flow as node are removed varies widely
- Indices with high top-*k* overlap impact the accommodated traffic similarly
- Highest resilience against the ECC-driven removals
- The locally-computed DC can approximate the global index with the worst impact over the maximum flow



Bren



Geant



Conclusions (1/2)

- A plethora of centrality indices have been proposed (since 1950)
    - Concept borrowed from Social Network Analysis and increasingly used in network protocol design
    - Our starting point : a novel classification scheme

- Correlation study between the seven most popular indices over ISP networks
    - Certain pairs found to be highly correlated
    - CC-HC, PG-DC (Expected)
    - BC-DC, ECC-CC (Not so trivial)
    - Top-5% overlap reveals more loose association
    - Warning: correlation is typically high but this is not uniform over the full ranking

Conclusions (2/2)

- Vulnerability of ISP networks to centrality-driven node attacks
- Network connectivity:
  - Centrality index pairs may exhibit dissimilar impact despite their high rank-correlation
  - ECC is consistently the index with the least impact
  - It is topology-dependent to approximate the (global) index with the worst impact using the locally-computed DC

- Traffic-carrying capacity:
  - Centrality pairs with high top-5% overlap impact the accommodated traffic in similar ways
  - DC can approximate closely the index with the most dramatic impact

Thank you!

Thank you!

Questions ?

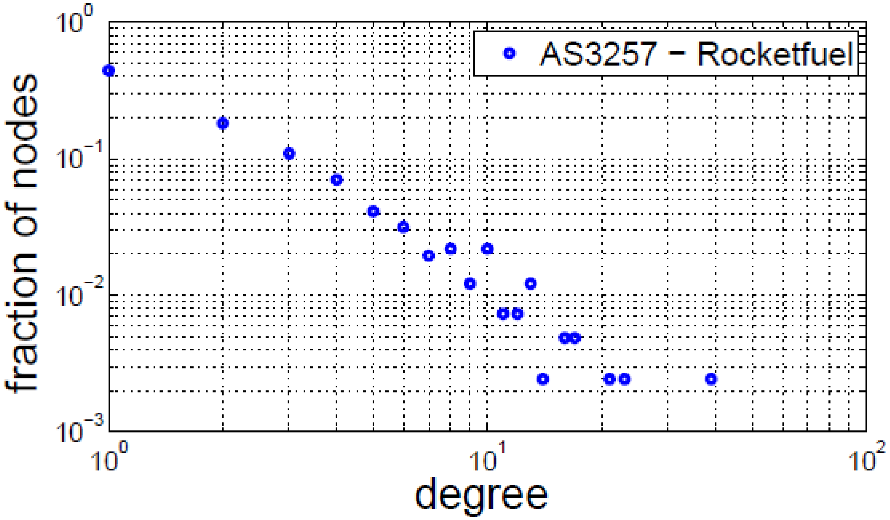
Back up slides

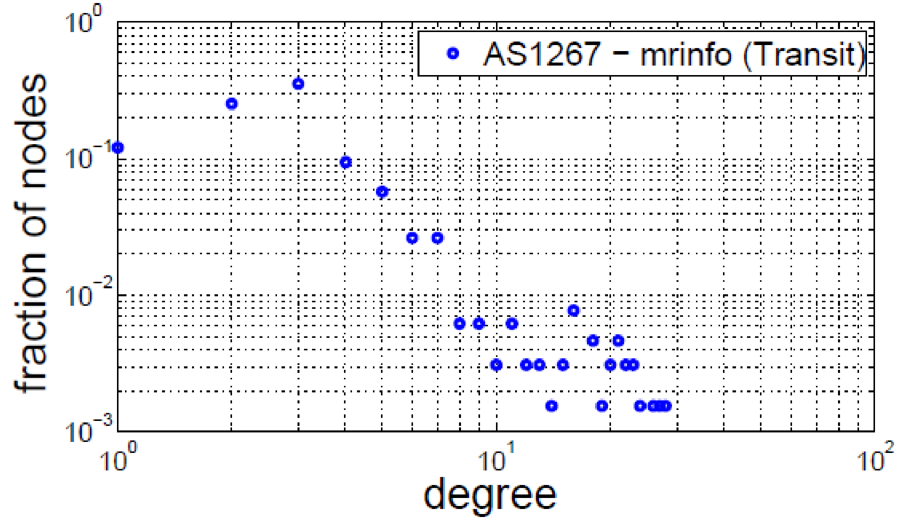
Definitions of seven popular centrality indices

- Degree Centrality $c_i^D = \frac{deg(i)}{N-1}$
- Betweenness $c_i^B = \frac{2}{(N-1)(N-2)} \sum_{j,k \in G, j \neq k \neq i} \frac{n_{jk}(i)}{n_{jk}}$
- Closeness $c_i^C = \frac{N-1}{\sum_{j \in G, j \neq i} d_{ij}}$
- Harmonic $c_i^H = \frac{1}{N-1} \sum_{j \in G, j \neq i} \frac{1}{d_{ij}}$
- Eigenvector $c_i^E = \frac{1}{\lambda} \sum_{j \in G} a_{ij} \cdot x_j$
- PageRank $c_i^{PR} = \frac{1-d}{N} + d \sum_{v \in B_i} \frac{R(v)}{L_v}$
- Eccentricity $c_i^{Ecc} = \frac{1}{max_{j \in V} d(i,j)}$

-N: the total number of nodes
-d_{ij}: geodesic path from i to j
-d: damping factor (equals 0.8)
-a_{ij} adjacency matrix
-λ, x eigenvalue & eigenvector respectively
-L_v out degree of page v

How does the degree distribution relate to correlation?





- High EC-CC correlation has been reported for synthetic scale-free graphs
- AS3257: Pearson r= 0.65 Spearman $\rho_v = 0.88$
- AS1267: Pearson r= 0.78 Spearman $\rho_v = 0.96$

- The actual association between two metric variants is not determined solely by the degree distribution