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

George Nomikos* Panagiotis Pantazopoulos* Merkourios Karaliopoulos† Ioannis Stavrakakis*
* Department of Informatics and Telecommunications
National & Kapodistrian University of Athens, Greece
Email: {gnomikos, ppantaz, ioannis}@di.uoa.gr
† Centre for Research and Technology-Hellas
Information Technologies Institute, Volos, Greece
Email: mkaraliopoulos@iti.gr

***Abstract*—The position of the nodes within a network topology largely determines the level of their involvement in various networking functions. Yet numerous *node centrality indices*, proposed to quantify how central individual nodes are in this respect, yield very different views of their relative significance. Our first contribution is then an exhaustive survey and categorization of centrality indices along several attributes including the type of information (local *vs.* global) and processing complexity required for their computation.**
We next study the seven most popular of those indices in the context of Internet vulnerability to address issues that remain underexplored in literature so far. First, we carry out a correlation study to assess the consistency of the node rankings those indices generate over ISP router-level topologies. For each pair of indices, we compute the full ranking correlation, which is the standard choice in literature, and the percentage overlap between the k top nodes. Then, we let these rankings guide the removal of highly central nodes and assess the impact on *both* the connectivity properties and traffic-carrying capacity of the network. Our results confirm that the top-k overlap predicts the comparative impact of indices on the network vulnerability better than the full-ranking correlation. Importantly, the locally computed degree centrality index approximates closely the global indices with the most dramatic impact on the traffic-carrying capacity; whereas, its approximative power in terms of connectivity is more topology-dependent.

I. Introduction

Social Network Analysis (SNA) provides a highly interdisciplinary theoretical framework for processing social information and analyzing social structures [1]. It draws heavily on graph models mapping individual actors within the social network to the graph vertices and their relationships to the graph (weighted) edges. It then leverages graph-theoretic concepts, metrics and results to answer questions about the relative importance of actors and the way information flows across it.

Centrality is one such concept/metric. To the best of our knowledge, it dates back to the work of Bavelas [2], who first gave a formal definition of node centrality in connected graphs as the sum of its geodesics (shortest-paths) to all other nodes. By that time significant sociological research was directed to the area of professional networks addressing how the position and power of individual actors relate to their social interconnections with the rest of the network and motivating a large research thread in the area of centrality indices. New indices (*e.g.*, [3]) were proposed and existing ones were adapted to apply to a broader range of scenarios [4]. The vast majority of work was heuristic and only a few of them attempted to come up with axiomatic definitions of centrality indices and the properties they should satisfy [5]. The highly-cited work of Freeman in [6] appears to have served as a turning point for this first wave of work. He reviewed a number of centrality indices and promoted three of them, *i.e.*, the closeness, degree, and betweenness, as the most representative ones. About the same time Bonacich had established the eigenvector centrality as a fourth, distinctly different but equally popular, index [7].

The research interest in the centrality concept revived in late 90's and early 2000, primarily through the works of physicists, who used centrality indices to explore the vulnerability and community structure of general network instances. Since then, centrality has found application to a broader set of disciplines beyond sociology. In computer science, in particular, insights from centrality indices are primarily exploited in the design of effective protocols for data networks (*e.g.*, [8], [9]). The trend is only catalyzed by the broader expectations about the evolution of a *Network Science*, which could serve as a basis for a unified treatment of all network types.

Motivation and objectives: Our main objective in this study is *to quantify how much information is embedded in centrality indices about the relative importance of Internet nodes for different network operations*. Given that the all centrality formulations proposed in literature are heuristic, the questions that naturally arise are how do these formulations compare in their assessments/predictions about the nodes' relative importance and which one(s) may be the "right one(s)" to consider as reference for more reliable predictions of network vulnerability.

The paper seeks to *systematically* address these questions by undertaking a three-step study with various instances of methodological innovation. The first step, which herein is briefly presented, involves a thorough survey and novel classification of the variety of centrality indices proposed in literature over the last sixty years. This classification is then used to select the seven most popular and representative indices for carrying out the two experimental steps of the study. Hence, as a second step, we derive the node rankings these indices induce over more than 40 *router-level* snapshots of network topologies and study their correlation. The correlation strength is assessed by the mainstream rank/linear correlation coefficients but also less widespread measures such as the percentage overlap in the lists of the k most central nodes. Finally, we compare

This work has been partially supported by EU-FP7 NoE EINS under grant agreement No 288021 and the University of Athens (ELKE-10812).

the seven indices with respect to their capacity to reveal the network vulnerability to node removals. Hence, we let the indices dictate the most central nodes to-be-removed and assess how the *network connectivity* properties but also its *traffic-carrying capacity* are affected.

Our results identify certain index pairs with consistently high full rank correlation across all datasets we experiment with. However, they also warn against careless interpretations of its values since a significant part of this correlation is shown to be due to nodes at the bottom of the rankings. On the other hand, the percentage overlap of the k most central nodes for the same pairs assumes clearly smaller values but predicts more accurately how similar is the impact on the network performance when two different indices are used to identify and remove the most central nodes. Notably when the node removals are driven by the single index that can be computed through local-only information (*i.e.*, Degree Centrality), the impact on the network traffic serving capacity approximates closely the maximum over the seven indices. The hint is that the added complexity of global indices may be circumvented while still exposing efficiently the network vulnerability to node removals.

The remainder of the paper is structured as follows: In Section II, we summarize a survey and the innovative classification scheme we have designed in [10] for numerous centrality indices proposed over the last decades. The correlation study of the seven most popular and representative indices is presented in Section III. Then, in Section IV we let those indices drive node removals over router-level network topologies and experimentally assess their impact on both their connectivity a properties and traffic-serving capacity. We discuss the related literature in Section V and conclude the paper with a summary of the main messages out of our study in VI.

II. A NOVEL CLASSIFICATION OF CENTRALITY INDICES

In this Section we briefly present the way we have characterized and classified the rich variety of centrality indices that we have run across in our study of the highly interdisciplinary 60-year-old literature. The interested reader is referred to [10] for a detailed description of the indices and the context within which they were originally proposed.

At a first-level the reviewed indices are divided into *node* (point) centrality and *graph* centrality indices. The former are addressed by the vast majority of the literature and concern individual nodes; whereas the latter are derived for whole graphs as functions of the individual node centrality indices. Then, node centrality indices are further characterized using three fundamental attributes, briefly discussed next:

Centrality context: topological vs. flow-aware. The vast majority of centrality indices takes only the network topology into account. They reflect either the distance of a node from all other network nodes [11] or the extent to which a node lies on *paths* connecting other network nodes [6]. Both types can be further differentiated as to whether they account only for geodesics between node pairs or a broader set of paths. Topological centrality indices also include the so-called spectral indices, which depend on the eigenstructure of a matrix (*e.g.*, adjacency or Laplacian) related to the network in question. The second set groups indices which attempt to factor the (predicted) network traffic in the centrality computation [12].

Underlying graph types. Most of the indices are defined over *connected, undirected, binary, static* graphs. Efforts to relax in turn each one of these four graph attributes have resulted in a plethora of indices that can cope with disconnected [13], directed [14], weighted and dynamic types of graphs [10].

TABLE I. THE NOTION THAT EACH INDEX REFLECTS FOR NODE i

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.

Computational Aspects. Centrality indices can be separated into local and global ones, depending on the extent of topological information that is required to compute them. Degree Centrality is clearly a local index while Betweenness and Closeness are global since they rely on network-wide geodesic paths [6]. To limit the scope of centrality computations one may use the sociological notion of the ego-network [15] or control the length k of the considered paths [16]. The corresponding computational complexity is of particular interest when centrality indices are embedded in network protocols.

Selecting centrality indices for experimentation. Out of the numerous indices reviewed in [10], we select the seven most popular ones that appear repeatedly in the literature, and, at the same time, capture a wide range of different notions of centrality. Those indices are the Degree (DC) [6], Betweenness (BC) [6], Closeness (CC) [6], Eigenvector (EC) [7], Harmonic Centrality (HC) [13], Pagerank (PG, with d=0.85 as typically used in literature) [14] and Eccentricity (ECC) [11]. Table I highlights the intuition behind each of the considered indices while Table II presents their formal (normalized) definitions and characterizes them according to the aforementioned classification attributes for a graph $G = (V, E)$ of $|V|$ nodes and $|E|$ edges.

III. CORRELATION STUDY OF CENTRALITY INDICES

In almost all instances, where centrality indices inform network protocols, what matters is the *ranking* of nodes induced by those indices rather than their absolute values. These rankings are subsequently used in the decisions made by the respective protocols. For example, in [8], the rankings determine whether a Delay Tolerant Network (DTN) node will forward a message to another DTN node it encounters; in [9], whether a content item will be cached at a Information-Centric Networking (ICN) node or not; and in [17] whether to search for a file in a given unstructured Peer-to-Peer (P2P) node or not. Likewise, in vulnerability analysis of the service migration protocol in [12], it is the *set* of the k, $k < |V|$ most highly-ranked nodes that matters, irrespective of their actual centrality values. The question that plausibly arises in every case is how similar are the rankings generated by each centrality index.

In this section, we carry out a thorough correlation study of these rankings, computed over a broad set of ISP router-level topologies. First, we calculate for each topology and node

TABLE II. PROPERTIES OF SEVEN POPULAR CENTRALITY INDICES UNDER A NOVEL CLASSIFICATION SCHEME

Centrality Index	Context: Topology aware			Context: Flow aware	Type of underlying graph: Binary/weighted		Directed/Undirected		Dynamic	Connected/Disconnected		Computational aspects: Information (local/global)		Complexity	Definition
	path	distance	spectral		B	W	D	U	D	C	D	L	G		
Betweenness (BC)	✓				✓	✓	✓	✓	✓	✓	✓		✓	$O(VE)$	$c_i^{BC} = \frac{2}{(N-1)(N-2)} \sum_{j \neq k \neq i} \frac{d_{j,k}(i)}{d_{j,k}}$
Closeness (CC)		✓			✓	✓	✓	✓	✓	✓			✓	$O(V(logV)E)$	$c_i^{CC} = \frac{N-1}{\sum_{j \in G, j \neq i} d_{i,j}}$
Degree (DC)	✓				✓	✓	✓	✓	✓	✓	✓	✓		$O(V^2)$	$c_i^{DC} = \frac{deg(i)}{N-1}$
Eccentricity (ECC)		✓			✓	✓	✓	✓		✓			✓	$O(V(logV)E)$	$c_i^{ECC} = \frac{1}{max_{j \in V} d_{i,j}}$
Eigenvector (EC)			✓		✓	✓		✓		✓			✓	$O(V^3)$	$c_i^{EC} = \frac{1}{\lambda} \sum_{j \in G} \alpha_{i,j} \cdot c_j^{EC}$
Harmonic (HC)		✓			✓	✓	✓	✓		✓	✓		✓	$O(V(logV)E)$	$c_i^{HC} = \frac{1}{N-1} \sum_{j \in G, j \neq i} \frac{1}{d_{i,j}}$
PageRank (PG)			✓		✓		✓			✓	✓		✓	$\Omega(\frac{E^2}{ln(1/(1-d))})$	$c_i^{PG} = \frac{1-d}{N} + d \sum_{v \in B_i} \frac{c_v^{PG}}{L_v}$

N: Total number of nodes, $d_{i,j}$: Shortest path length from i to j, $d_{j,k}(i)$: Shortest path length via k, $\alpha_{i,j}$: Adjacency matrix element d : Damping factor, B_i: Set of nodes linked to i, L_v: Out-degree of node v

in it the seven centrality indices (Section II), thus generating seven different node rankings per topology. Then, we compute pairwise correlation measures over these rankings. We consider two different measures, one accounting for the full node rankings and the other only for the most highly-ranked nodes.

A. Index correlation measures and router-level topologies

Index correlation measures. The first correlation measure is the nonparametric Spearman's rank-correlation coefficient, ρ_V, and is computed over the full node rankings. The coefficient assesses how well a monotonic function can describe the rankings induced by the two centrality indices on the network nodes. For a given network topology node set V, it is:

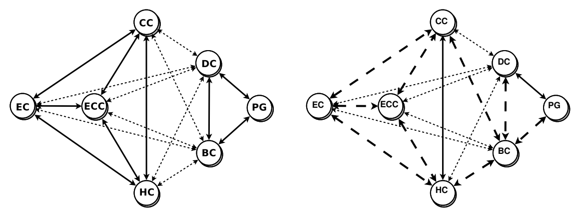
$$\rho_V(C_1, C_2) = 1 - \frac{6 \sum_{u \in V} (r_{C_1}(u) - r_{C_2}(u))^2}{|V|(|V|^2 - 1)}$$

where $r_{C_1}(u)$ and $r_{C_2}(u)$ are the ranks of node u in line with centrality indices C_1 and C_2, respectively. It lies in $[-1, 1]$, with high positive (negative) values denoting strong positive (negative) correlation[1]. The second correlation measure is the percentage overlap between the sets of the k most highly ranked (top-k) nodes that are generated by two indices.

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

Contrary to the Spearman's ρ_V, the percentage overlap is computed over a subset of the full node rankings and takes values in $[0, 100]$. The relevance of the two measures depends on the usage context of centrality-based ranks. The decisions that relate to the DTN forwarding, CCN caching and P2P node search examples rely on full node rankings; whereas, vulnerability analysis is usually concerned with the subset of nodes that are important ("central") for the network. High correlation between the rankings of two indices implies that a computationally complex index can be approximated by a simpler one without significant penalties for the protocol operation or the conclusions of the vulnerability analysis.

[1]We have also computed two other popular coefficients; one is the Kendall's τ for rank-correlation which generally gives similar results with ρ. The other is the Pearson coefficient r that assesses how linear is the relationship between the *actual values* of the indices rather than their rankings, is found high for highly rank-correlated centrality pairs, yet of lower strength [10].



a. b.

Fig. 1. Graph-based illustration for the averages values of the Spearman coefficients (a) and top-5% overlap (b) among centrality indices. In (a) solid bold and dashed plain lines denote coefficients in the intervals [0.7-1],[0.3-0.7), respectively. In (b), solid bold, dashed bold and dashed plain lines denote overlap higher than 70%, between 40-70%, and lower than 40%, respectively. Over CAIDA, BC-CC and BC-HC are exceptions to this rule exhibiting reduced top-5% overlap values.

Router-level ISP topologies. All our experiments are carried out over four datasets. Three of them relate to measurement projects and are referred to as Rocketfuel [18], CAIDA [19], and mrinfo (Tier-1 and Transit) [20] datasets, respectively. They report *binary* router-level graphs for different Internet ASes. We experiment with 9 RocketFuel snapshots, 7 CAIDA, 8 Tier-1, 6 Transit with size range 41-9418, 1831-81121, 76-741, 336-1240 nodes, respectively. On the contrary, the last dataset, called Topology Zoo (we use 18 snapshots of 20-74 nodes), contains *capacitated* topologies at the router- and Point-of-Presence (PoP) level [21], collected directly by network operators of academic and research networks. The basic properties of all datasets are summarized in [10].

B. Results

Full-ranking correlation over binary graphs. Table III reports averages of the Spearman's rank correlation, as computed over all snapshots of a given binary dataset (*i.e.*, CAIDA, Rocketfuel, MrInfo -Tier1 and -Transit). The first remark is that not a single centrality pair is negatively correlated over any of the studied topologies. We empirically characterize the pairwise index correlation as *high* and *low* when the corresponding ρ_V values lie in the intervals [0.7,1] and [0.3,0.7), respectively. On the other hand, two indices are considered non-correlated when their ρ_V lies in [0-0.3). The second point is that the indices' correlation values follow similar trends across all datasets so that they can summarized graphically in a graph like the one of

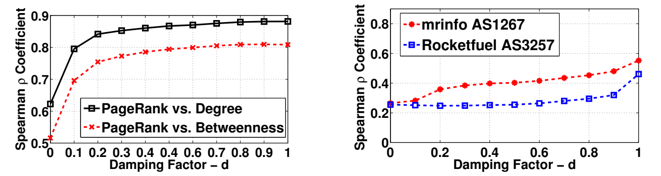
Fig. 1.a. Solid bold edges in the graph denote high correlation between two indices, whereas dashed edges represent low values. No edges are added for non-correlated index pairs. In what follows, index pairs of interest are discussed in detail. Where appropriate we refer to studies reporting relevant results on different kinds of networks. The full set of results can be found in [10].

TABLE III. SPEARMAN COEFFICIENT AVERAGES FOR ALL DATASETS

	CC	HC	EC	ECC	DC	BC	PG	dataset
CC	1							CAIDA
	1							RocketFuel
	1							MrInfo-Tier1
	1							MrInfo-Transit
HC	0.99	1						-//-
	0.98	1						
	0.95	1						
	0.99	1						
EC	0.93	0.95	1					-//-
	0.80	0.83	1					
	0.66	0.69	1					
	0.86	0.88	1					
ECC	0.84	0.84	0.84	1				-//-
	0.73	0.67	0.56	1				
	0.80	0.69	0.52	1				
	0.89	0.88	0.75	1				
DC	0.28	0.28	0.29	0.25	1			-//-
	0.48	0.53	0.45	0.38	1			
	0.43	0.59	0.47	0.30	1			
	0.50	0.55	0.49	0.45	1			
BC	0.29	0.29	0.28	0.27	0.90	1		-//-
	0.45	0.50	0.40	0.37	0.94	1		
	0.50	0.61	0.30	0.38	0.69	1		
	0.54	0.58	0.48	0.47	0.88	1		
PG	0.04	0.05	0.05	0.04	0.86	0.80	1	-//-
	0.25	0.30	0.14	0.20	0.83	0.80	1	
	0.34	0.49	0.30	0.24	0.90	0.74	1	
	0.40	0.44	0.36	0.35	0.92	0.88	1	

Betweenness vs. Degree centrality: Degree centrality (DC) captures, at least phenomenally, a completely different notion of centrality than Betweenness (BC). DC takes into account only the node's local neighbors, whereas BC considers the position of the node within the whole network. Therefore, in some cases DC can evaluate nodes' position very differently than BC; it may overestimate the importance of nodes belonging to isolated subgraphs (high DC-low BC) or underestimate the role of nodes acting as bridges between groups of nodes (low DC-high BC). On the other hand, high-degree nodes have better chances to be parts of the shortest paths linking node pairs. In our datasets, the two indices are found consistently highly correlated, in agreement with earlier studies [15], [22], [23] that report positive *Pearson* correlation between DC and BC over a wide range of networks such as random graphs and real-world complex networks.

Pagerank vs. Degree centrality: Another persistent result, immediately apparent from Figure 1.a, is the strong correlation between Pagerank (PG) and DC. Pagerank is principally defined for digraphs discriminating between incoming and outgoing connections at each node. The DC-PG correlation increases with the damping factor d of PG, as shown in Fig. 2.a. It also shows similar association between the d factor and the PG-BC ρ values. Taking into account the aforementioned strong BC-DC correlation, a triangle-like schema emerges and may be of practical importance as it relates DC, the only local, index with two globally-determined ones. Grolmusz shows in [24] for undirected general graphs that Pagerank is statistically close but not identical to the degree distribution. Positive correlation between the three indices (PG-DC-BC), with ρ values in [0.66, 0.95] for all three index pairs, is also reported in [25] over coauthorship real-world data (directed graphs).



a. AS1299 b.

Fig. 2. a) Rank-correlation scaling as PG increasingly depends on DC and BC. b) Rank-correlation between EC and PG against the damping factor d.

Pagerank vs. Eigenvector centrality: PG, and EC centrality are the two spectral indices we experiment with. Both express the stationary probability of a random surfer to reside on some page while moving on the Web graph. Hence, one would expect some positive correlation between these indices. However, our results indicate the absence of such a relationship. A possible cause is that their actual interpretation differs as, contrary to EC, the PG centrality utilizes the damping factor d to determine the "jump" probability. However, as a couple of indicative experiments suggest (Fig. 2.b), the rank correlation between the two metrics increases yet does not reach very high values as d moves to unity *i.e.*, the surfer moves only to neighboring pages. It seems then that d can only partially justify the poor PG-EC correlation strength; as the PG formula suggests (Table II) a node's (*i.e.*, Web page) PG rank value is evenly divided (L_u term) over its neighbors, which for the case of undirected graphs corresponds to its DC value. The fact that DC index is found to be weakly correlated with EC (Table III) can further distort any anticipated PG-EC correlation.

Eccentricity vs. Closeness centrality: Another strong correlation that we observe in our study is between the Eccentricity and Closeness centrality indices. Recalling the definitions of the two indices (ref. Table II), there is absolute positive ECC-CC correlation if it holds that $ECC(n_1) > ECC(n_2)$ whenever $CC(n_1) > CC(n_2)$, for all $n_1, n_2 \in V$. We can rewrite the former equation as $max_{j \in V} d_{n_2,j} > max_{j \in V} d_{n_1,j}$ and the latter as $\sum_{j \in V} d_{n_2,j} > \sum_{j \in V} d_{n_1,j}$. Hence, looking at the last two inequalities, the question becomes when the order in maximum index values is also preserved for their averages over the studied graph. This holds in several trivial graphs (*e.g.*, line graph, rectangular grid) but not in all graphs. One simple counterexample is the 4-node star network with a 2-node line graph attached to one of its leaf nodes (compare the two indices for the hub node and the leaf node, where the line is attached).

Additional remarks: There exist further centrality pairs yielding positive correlations, which are less straightforward to reason about. For instance, in our results, high rank correlation has been observed for pairs such as Eigenvector-Harmonic and Eigenvector-Closeness centrality. These findings seem consistent with previous results. Iyer *et al.* [26] have noticed that synthetic scale-free networks (whose degree distribution follows a power law, at least asymptotically) present moderate positive Pearson CC-EC correlation. Higher values (r=0.61) are reported for networks with exponential degree distribution. We have tried to identify how the degree distribution relates to the EC-CC correlation. In Figure 3 left, we plot in log-log scale the degree distribution of a 411-node large AS out of the RocketFuel dataset, as a representative sample, with positive

Pearson EC-CC correlation (r=0.65). The straight-line points to power-law degree distribution suggesting that this may be beneficial for the positive correlation, as in [26]. On the other hand, the scale-free property is not a necessary condition for high EC-CC correlation. In Figure 3 right, the degree distribution of a 645-node large mrinfo Transit AS clearly deviates from a power-law pattern, yet it features a considerably higher Pearson coefficient (r=0.78). Similar remarks hold for the EC-CC rank correlation over these snapshots (Spearman ρ_V=0.88 and ρ_V=0.96, respectively).

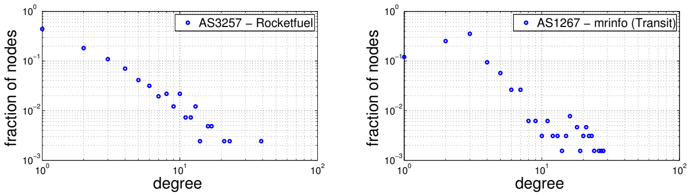


Fig. 3. Degree distribution for two indicative snapshots.

Top-k percentage overlap over binary graphs. So far, our correlation analysis has taken into account the full rankings produced by the seven centrality indices. We now focus our attention on the top-5% most central nodes identified by each index and investigate how large are the overlaps between different rankings. The motivation for this set of experiments is the existence of network protocols that seek to exploit a small set of the most central nodes [12]. Likewise, vulnerability studies of Internet graphs, as the one we carry out in Section IV, are concerned with such node subsets.

In Figure 1.b we show a summarizing graph-based illustration of the overlap scores among the seven centrality indices. The bold solid lines (*e.g.*, between CC-HC) denote high top-5% overlap between two indices *i.e.*, beyond 70%. The dashed solid lines (*e.g.*, between EC-HC) reflect overlap values between 40-70%, whereas the dashed plain lines represent looser relations. Additionally, figure 4 presents the average overlap of nodes over all ASes of each dataset for the most significant centrality pairs. On the one hand the overlap of some indices (*e.g.*, BC-CC or HC-BC) appear to be highly sensitive to the considered topology, with differences that reach 40% across different datasets. On the other, all pairs found earlier to be strongly correlated in terms of full rankings, appear to be more weakly associated in terms of overlap values[2]. Exceptions to that rule are the HC-BC and CC-BC pairs that represent a slight increase of the relation strength when passing from the rank correlation to the overlap measure. Overall, only two of the centrality index pairs combine high overlap values with strong full rank-correlation (see Fig. 1.a). PG-DC and HC-CC, both exhibiting larger than 80% overlap in the top-5% node rankings they induce across all datasets, whereas all the other pairs hardly exceed the 60% value. This result should come as no surprise since rank correlation is determined over all network nodes rather than a subset of cardinality k.

Let us look closer into the BC-DC pair. Figure 5 illustrates how the number of nodes with DC=1 affects the rank correlation coefficient. It seems that (especially for Caida and RocketFuel datasets) the Spearman values between the two indices increase with the number of DC=1 nodes. These nodes are expected to positively contribute to the DC-BC correlation as they also exhibit the lowest-ranked betweenness value (*i.e.*, BC=0). At the same time, the ones with the top BC and DC values may not necessarily coincide as indicated in Table IV. The above results suggest that the high DC-BC correlation is mainly due to nodes of lowest ranks. This observation warns against the actual value of high Spearman rank correlation coefficients between two indices. On the other hand, the overlap measure does not suffer from similar biases. The repercussions of this will become clearer in the results of the Section IV experiments.

[2]The characterization retains a loose empirical meaning since comparing a correlation coefficient with the % overlap value is not straightforward.

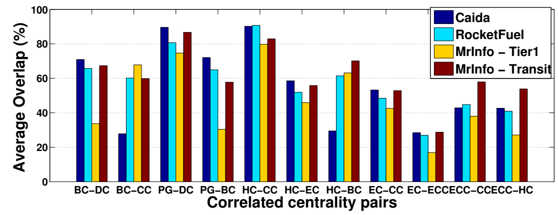


Fig. 4. Mean overlap (%) between the top-5% nodes of centrality rankings.

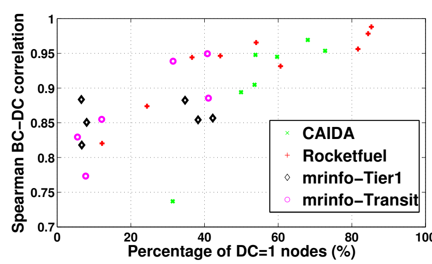


Fig. 5. The relation between the BC-DC rank correlation and the percentage of nodes with degree equal to one.

TABLE IV. RANK CORRELATION VS. OVERLAP(%) BETWEEN BC-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%

Correlation/overlap results over capacitated graphs. We have carried out a brief correlation study of the considered indices over the topology Zoo dataset. To determine the node rankings we had to carry out the centrality indices computations over weighted graphs. This was mainly a question of computing shortest paths over weighted graphs. Regarding the spectral indices, in the Topology Zoo experiments we only employ the EC index that lends to a straightforward extension over the weighted graphs [10]. As such, we compute the Spearman coefficients for the centrality pairs across all 18 snapshots and present the average and variance values in Table V. Those index pairs that were measured earlier to be strongly correlated over the binary graphs (Figure 1.a), generally maintain similar relations over the capacitated Zoo networks. BC-DC correlation is again found high yet not as close to unity as before; ECC and EC appear in most cases highly correlated except for a few topologies that contribute to a high variance value for the coefficient average. Clearly, these results are shaped by both the topology and the link capacity values that are now taken into account for the corresponding index computations.

TABLE V. TOPOLOGY ZOO: SPEARMAN AVERAGES AND VARIANCE

	BC	CC	DC	EC	HC	ECC
BC	1					
CC	0.68±0.01	1				
DC	0.75±0.01	0.85±0.02	1			
EC	0.59±0.03	0.94±0.01	0.79±0.04	1		
HC	0.68±0.01	0.98±0.01	0.87±0.01	0.95±0.01	1	
ECC	0.64±0.01	0.87±0.04	0.80±0.02	0.77±0.16	0.86±0.04	1

TABLE VI. TOPOLOGY ZOO: AVERAGE TOP-15% OVERLAP (%)

	BC	CC	DC	EC	HC	ECC
BC	1					
CC	75.55	1				
DC	78.38	84.36	1			
EC	67.40	84.07	78.06	1		
HC	75.55	89.63	85.10	87.77	1	
ECC	69.99	77.91	73.62	71.24	73.09	1

In Table VI we present our results for the overlap between the $k\%$ top central nodes averaged over the whole Zoo dataset. As their size is relatively small, we set k=15% in order to each time avail vectors of at least 5 nodes' size. Compared to the top-k overlap measured over the binary graphs, we have found the same index pairs to exhibit high values; one exception is the HC-DC pair which now appears of considerably high overlap. In the next section we will see how theses overlap values reflect on the (similar) effects of the corresponding node removals.

IV. CENTRALITY AND NETWORK VULNERABILITY

The correlation study yields a first indirect indication of how different centrality indices compare and whether they could be interchanged in the context of a network protocol or analysis that draws on node rankings. The ultimate reply to this question is, however, protocol/analysis-dependent. In this section, we seek to come up with a reply in the context of the network vulnerability to node failures. More specifically, we ask how similar are the conclusions about the network vulnerability when relying on different centrality indices to identify and remove its most central nodes[3].

The network vulnerability analysis is of interest to various parties. A potential attacker would like to know which index results in node removals with the most significant impact on the network performance so as to orchestrate the most effective attack. From the network operator's side, the dual aim is to identify and better protect those critical nodes, whose failure would result in maximum network performance degradation. In this paper, we relate the term "performance" to fundamental *connectivity* and *traffic capacity* properties of the network rather than the scores achieved by specific protocols/applications. This way we get away with their engineering details that shape the end impact and place the emphasis on the network topologies per se.

[3]In this paper, nodes are removed simultaneously after being ranked in order of decreasing centrality. An alternative called *sequential* targeted attack strategy, is to recalculate the rankings of the residual nodes after each removal. As intuitively expected and shown in [26], [27] the impact of such sequential node removals upon the network connectivity is more dramatic. Expanding our study to the sequential node removal case is straightforward.

A. Centrality-driven node removals and connectivity

The experiments of this section explore how the size of the giant connected component and the total number of connected components in each topology are affected when up to 5% of the network nodes are removed[4]. The experiments are carried out over the binary datasets described in the subsection III-A.

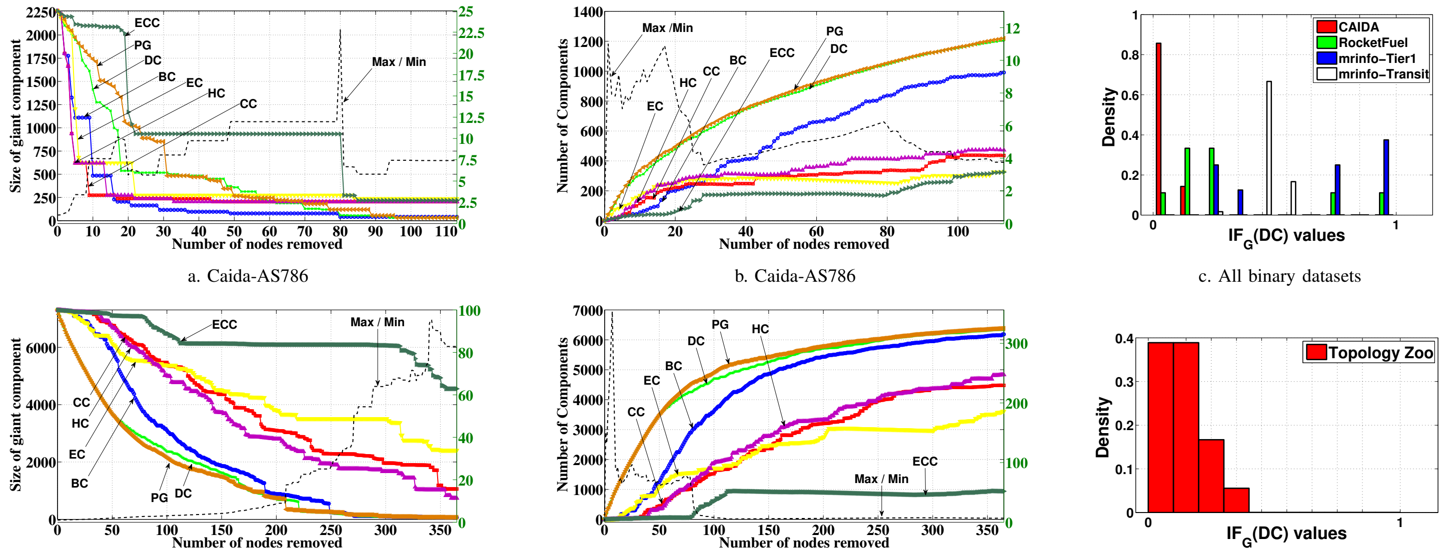
Figures 6a,b-d,e show representative plots of the two connectivity measures as a function of the number of removed nodes. Besides the two connectivity measures also plotted in dashed line (and measured on right Y axis) is the Max/Min ratio. This is the ratio of the maximum over the minimum value the connectivity measure assumes for a given number of removed nodes and over all centrality indices. The Max/Min ratio essentially quantifies the variance in the connectivity measure value as a result of the choice of centrality index for identifying central nodes. Next, we comment on the experimentation outcomes and relate them to the earlier observed correlations.

Size of giant connected component (GCC): The GCC size reflects the number of nodes that can communicate with each other. Figs. 6.a,d suggest that removing those vertices that the ECC index identifies as most central has the minimum impact on GCC. All other indices expose more quickly the vulnerability of the network but we cannot identify any dominance relationship among them that persists over all datasets. However, the behavior of certain index pairs such as the HC-CC and PG-DC is in good agreement with the earlier observed strong correlation, both full rank and top-k percentage overlap (ρ_V and ov_V exceed 0.85 and 85%, respectively); indeed, the corresponding curves in Figs 6.a,d (partially) coincide or exhibit small GCC size differences with the number of removed nodes. A closer look reveals that it is the top-k overlap between two indices, rather than their rank-correlation, that essentially determines how similar is the impact of the corresponding removals. A relevant example is the BC-DC pair over AS1239. Their highly dissimilar impact, as shown in Fig 6.d, can be predicted by the 68% top-5% percentage overlap; it cannot be inferred by the high Spearman rank correlation (ρ_V=0.94).

In [27] Holme *et al.* show that DC- and BC-driven node removals are equally harmful over synthetic (scale-free) graphs, while a distinct real-world co-authorship network appears more vulnerable to BC-driven attacks. In our broad dataset of real-world topologies we do not witness the latter effect; on the contrary, the local DC index occasionally has more dramatic impact than the global BC. Overall, a concluding note would be that *any two indices measured with high top-k overlap values are expected to give rise to similar GCC sizes, and vice versa. The full-rank correlation values are not always in line with the experienced impact due to the biases discussed in Section III.*

Number of connected components: Again, the ECC index yields node removals that result in minimum network fragmentation (Figs. 6.b,e). According to the ECC definition [11], a node is central when its maximum distance to any other node is close to the radius of the graph. Hence, a node can exhibit a significantly low ECC value when only a few other nodes lie

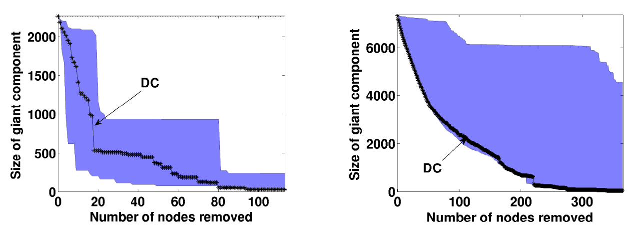
[4]The measured impact is essentially a worst-case result for the network connectivity. In reality, some hidden redundancy remains unnoticed even if our datasets have been extracted by different topology discovery tools.



a. Caida-AS786 b. Caida-AS786 c. All binary datasets

d. Rocketfuel-AS1239 e. Rocketfuel-AS1239 f. Capacitated dataset

Fig. 6. a,b,d,e) Effect of node removals on the size of the giant-connected component (a,d) and the number of components (b,e) for two indicative ASes. c,f) Empirical probability mass function of the $IF_G(DC)$ measured *w.r.t* the size of the giant component (c) and the max flow accommodated by Topology Zoo (f).



a. Caida-AS786 b. Rocketfuel-AS1239

Fig. 7. Envelope plots of the DC-based node removal effects on GCC.

far away (from it) in the topology. This sensitivity makes ECC assign less significance to nodes considered highly-central according to other indices. So the top-5% ranked nodes may not actually include those holding prominent network locations and this prevents the fast fragmentation of the topology. In sharp contrast, DC and PG are together dominant in terms of partitioning the topology, as their high correlation suggests. Interestingly, DC, a purely local index succeeds in removing nodes that play critical role in connectivity as opposed to the other global and more complex ones (except PG). On the other hand, BC and DC which were also found strongly rank-correlated yet of weaker top-k overlap, have different impact on the connected components. Removing nodes according to DC, the number of components increases constantly compared to the impact of BC. This implies that the network connectivity mainly relies on strategic hub-nodes rather than bridging nodes that are typically of high BC.

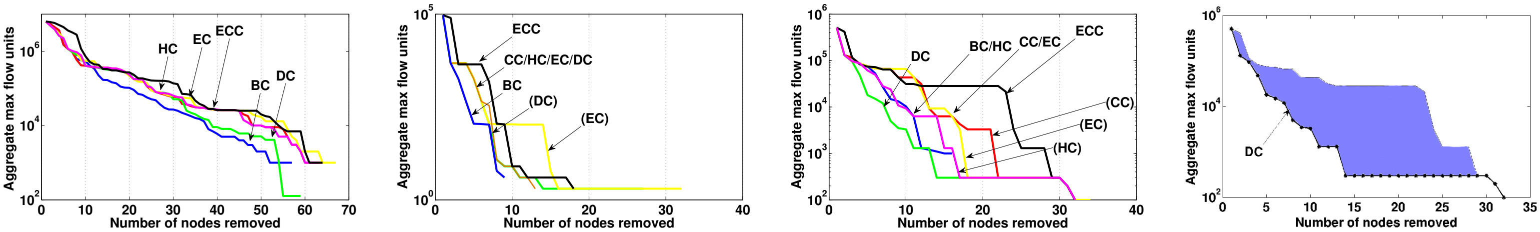
Local vs. global centrality indices: Figures 6.a,b,d,e clearly show that the removal of the most central nodes affects differently the connectivity measures depending on which centrality index is used to determine them. For each number k of removed nodes, one can identify best- and worst-case values, $m_{bc}(k)$ and $m_{wc}(k)$ respectively, for the two performance metrics. These values may be obtained by different centrality indices as the considered metric m changes and outline an envelope, marked by the shaded area in Fig. 7. What we ask next is where in this envelope the metric values corresponding to the degree centrality, lie. Essentially, we seek to quantify how close to the best-/worst-case is the impact of removals when directed by the single locally computable centrality index. To this end, for each centrality index c, topology G, number of removed nodes k and performance metric $m(k;c)$ we define a normalized distance measure, hereafter called impact factor $IF_G(k;c)$ as:

$$IF_G(k;c) = \frac{|m(k;c) - m_{wc}(k)|}{|m_{bc}(k) - m_{wc}(k)|}$$

Note that depending on the metric, the worst-case value may coincide with the minimum or maximum value the metric gets over all indices. It is then straightforward to derive a topology-average measure of the impact factor as:

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

where $\mathcal{K}$ is the set of k values considered in the evaluation. Clearly, both $IF_G(k;c)$, $k \in \mathcal{K}$ and $IF_G(c)$ take values in $[0,1]$. We are particularly interested in $IF_G(DC)$ and Fig. 6.c plots the empirical probability mass function of the $IF_G(DC)$ values over all topologies of a given dataset, when the metric m is the size of the GCC. Despite its local nature, DC-driven node removals in most cases affect significantly the GCC size. To which extent this impact approximates the worst-case over all indices depends on the underlying topology. Over the CAIDA networks DC closely approximates the low end of the envelope. The approximation is looser over Rocketfuel, whereas in the mrinfo (Tier-1) and (Transit) networks, considerable mass is accumulated at medium and high $IF_G(DC)$ values,



a. Uninett I_mean b. Carnet c. Bren d. Bren

Fig. 8. a,b,c) Impact of node removals on the maximum flow that Zoo topologies accommodate. When curves coincide, a single identifier points to multiple indices; when they become separated, each one is pointed with its index in parenthesis. d) Envelope plot of the DC-based node removal effects on the max flow.

respectively.

B. Centrality-driven node removals and traffic capacity

We now turn our attention to a much less investigated topic, the comparative impact of centrality-driven node removals on the network traffic serving capacity. Such a task is not straightforward. One approach would be to consider a given traffic matrix, determining either the node pairs that exchange traffic only or the node pairs plus the average traffic loads that are (expected to be) served for each one of them. Then, the traffic-serving capacity of the network could be given by the solution of an instance of the multicommodity flow (MCF) problem [10]. Yet doing so bears some significant challenges: First, the traffic demand matrix is rarely known a priori and often varies broadly over different time scales. This may be partially addressed by employing synthetically generated dynamic traffic matrices [28]. However, such a choice introduces an extra variable in the process of assessment *i.e.*, the accuracy of the traffic model. Therefore, it does not provide us with a solid reference for comparison of the metrics. Secondly, and most importantly, the MCF problem is an NP-complete problem [29], with the computational complexity raising fast with the number of commodities. To overcome those limitations, we have taken a simpler approach and estimate the traffic serving capacity of the network as the sum of maximum flows over all network node pairs. Namely, we iterate over all node pairs and for each pair we solve an instance of the maximum flow problem, *i.e.*, compute the maximum traffic load that can be served by the network when only the particular pair transfers traffic across the network. Clearly, this sum is a (very) loose upper bound of the traffic load that can simultaneously be served by the network. However, it provides a traffic load-neutral measure of what can the network carry and how is this affected when a variable number of nodes is removed. To solve the maximum flow problem we have used the Edmonds-Karp algorithm [30] with a $O(VE^2)$ polynomial-time complexity.

Experimentation methodology and results. Our experimental study is carried out over the Zoo Internet topologies with capacitated links, described in III-A. We remove nodes in decreasing order of centrality and measure the aggregate maximum flow over all node pairs. The computed aggregate maximum flow over an indicative set of networks is plotted in Figs. 8.a-c. We have obtained similar results for the rest of Zoo datasets (totally 18 snapshots). The rate of aggregate max flow reduction with the fraction of removed nodes varies wildly. This results in high best- to worst-case flow values and wide envelopes, as shown in Fig. 8.d. Highly correlated index pairs, especially those with high top-k percentage overlaps, affect the accommodated flow in similar ways (*i.e.*, intersection of corresponding curves). In particular, certain index pairs that have been earlier measured with high rank-correlation, and most notably top-k overlap, yield similar curves over a sequence of removals; for instance, the highly associated pairs of EC-CC, EC-HC and HC-CC (see Tables V and VI) are seen in Figs. 8.b,c. Similar impact of BC- and DC-driven node removals has been reported over synthetic graphs (*i.e.*, Erdős-Rényi and small-world networks) in [31]; in our case the impact of those indices is typically different over the Internet snapshots. Finally, weakly correlated pairs *e.g.*, EC-BC, ECC-BC, inline with intuition, yield well-separated flow curves.

On a positive note, when node removals are driven by the DC index, the resulting aggregate maximum flow in most cases of Fig. 8.a-c is very close to the worst achieved over all indices. This is more clearly shown in the empirical probability mass function of the $IF_G(DC)$ measure in Fig. 6.f, whose mass is highly concentrated at (very) low values close to zero. On the contrary, the considered networks exhibit their highest resilience against the ECC-driven node removals. This behavior can be explained along the same arguments employed earlier, when discussing how node removals affect the connected components. Having a single node *i.e.*, the furthest one, determine ECC may result in some of the most central nodes not being included in the top positions of the ECC ranking.

V. Related work

Regarding survey studies, Freeman back in 1979 reviewed several centrality indices proposed by that time and reduced them down to three fundamental notions, expressed by the degree, closeness and betweenness centrality [6]. Much later, Borgatti [32] introduced a typology of the different types of network flows and associated the various centrality measures with the flows that they are most appropriate for. A graph-theoretic review in [16] classifies centrality measures according to their computational requirements. Compared to them, we review a great body of centrality indices of different origins and classify them along multiple dimensions [10]. Particular care is taken for "engineering" properties such as the index computational complexity in distributed Internet environments.

With respect to centrality correlation studies, we are aware of two works that compute linear correlation between the DC and BC indices: one over a random network and a couple of real-world topologies with a single router-level snapshot [22] and another over three AS-level snapshots [23]. Neither of them assesses how the network is affected when different indices are used to direct node removals. Work along this thread typically addresses synthetic graphs and the removals' impact is measured through purely topological measures. Hence, in [33] the scale-free topologies are found vulnerable to the removal of high-degree nodes and in [27] removals of high-DC and -BC nodes in an AS-level topology are found equally harmful in terms of the inverse geodesic length and the number of connected components. More recently, Trajanovski *et al.*in [34] consider both random node failures and centrality-driven attacks in the context of a more general topological robustness framework. Experimenting with random graphs, power grids, railway and co-authorship networks the authors show that many centrality indices drive removals of similar impact and that DC and EC relate to the most harmful ones. Our study sets its focus on Internet *router-level* graphs and relates their vulnerability also to *non-topological* properties.

VI. Conclusions

We have undertaken a systematic approach to study the relevance of node centrality indices to the Internet vulnerability. Departing from an exhaustive survey and a novel classification scheme of numerous centrality indices, we have carried out a thorough correlation study of the node rankings the seven most popular indices generate over a broad set of ISP router-level topologies. Then, we have experimentally assessed the impact of node removals determined by those rankings. Contrary to previous works that consider only *network connectivity* issues we have extended the vulnerability context to the network *traffic-serving capacity*. Our main results follow:

- Certain index pairs (such as DC-BC, DC-PG) were consistently found to be high (rank-)correlated across all datasets. Yet a significant part of the high full rank correlation is due to the nodes that are ranked last (e.g., DC=1, BC=0) so that the association weakens when we measure the overlap between the sets of the top-5% most central ones.
- Node removals based on initial centrality rankings showed that index pairs may exhibit dissimilar impact on the connectivity despite their high (rank-)correlation. As expected, it is the top-5% overlap between the top nodes in those rankings that more precisely prescribes how different is the impact. This is a warning against the widespread use of full rank correlation as a proxy for the "equivalence" of two indices.
- ECC is consistently the index with the least impact. On the other hand, local-only information (DC) will be *practically* used to approximate the index with the *worst* impact. In terms of connectivity, such an approximation depends on the underlying network. In terms of the network traffic capacity, the approximation is highly effective implying that the complexity of global indices can then be escaped.

References

[1] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge Univ Pr, 1994.

[2] A. Bavelas, "A mathematical model of Group Structure," *Human Organizations*, vol. 7, pp. 16–30, 1948.

[3] L. Katz, "A new status index derived from sociometric data analysis," *Psychometrika*, vol. 18, pp. 34–43, 1953.

[4] M. Beauchamp, "An improved index of centrality," *Behavioral Science*, vol. 10, pp. 161–163, 1965.

[5] G. Sabidussi, "The centrality index of a graph," *Psychomatrika*, vol. 31, pp. 581–603, 1966.

[6] L. C. Freeman, "Centrality in social networks: Conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239.

[7] P. Bonacich, "Power and centrality: A family of measures," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[8] E. M. Daly and M. Haahr, "Social network analysis for information flow in disconnected delay-tolerant manets," *IEEE Trans. Mob. Comput.*, vol. 8, no. 5, pp. 606–621, 2009.

[9] W. K. Chai *et al.*, "Cache "less for more" in information-centric networks," in *Proc. of IFIP Networking*, Prague, Czech Rep., May 2012.

[10] G. Nomikos *et al.*, "The multiple instances of node centrality and their implications on the vulnerability of ISP networks," Tech. Rep., 2013. [Online]. Available: http://arxiv.org/abs/1312.4707

[11] P. Hagea and F. Harary, "Eccentricity and centrality in networks," *Social Networks*, vol. 17, no. 1, pp. 57–63, Jan. 1995.

[12] P. Pantazopoulos, M. Karaliopoulos, and I. Stavrakakis, "Centrality-driven scalable service migration," in *23rd International Teletraffic Congress (ITC'11)*, San Francisco, USA, Sept. 2011.

[13] Y. Rochat, "Closeness centrality extended to unconnected graphs: The harmonic centrality index," *In proc. of ASNA*, Zurich, Switzerland, 2009.

[14] L. Page *et al.*, "The pagerank citation ranking: Bringing order to the web," Stanford University, Technical Report, 1998.

[15] P. Pantazopoulos, M. Karaliopoulos, and I. Stavrakakis, "On the local approximations of node centrality in Internet router-level topologies," in *7th IFIP IWSOS, Palma de Mallorca, Spain*, 2013.

[16] S. Borgatti and M. Everett, "A Graph-theoretic perspective on centrality," *Social Networks*, vol. 28, no. 4, pp. 466–484, Oct. 2006.

[17] L. A. Adamic *et al.*, "Search in power-law networks," *Physical Review E*, vol. 64, no. 4, Sep. 2001.

[18] N. T. Spring *et al.*, "Measuring ISP topologies with rocketfuel." *IEEE/ACM Trans. Netw.*, vol. 12, no. 1, pp. 2–16, 2004.

[19] The CAIDA UCSD Internet Topology Data Kit ITDK2011-10. [Online]. Available: http://www.caida.org/data/active/internet-topology-data-kit/

[20] J.-J. Pansiot *et al.*, "Extracting intra-domain topology from mrinfo probing," in *Proc. PAM*, Zurich, Switzerland, April 2010.

[21] S. Knight *et al.*, "The internet topology zoo." *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1765–1775, 2011.

[22] C.-Y. Lee, "Correlations among centrality measures in complex networks." [Online]. Available: http://arxiv.org/abs/physics/0605220

[23] A. Vázquez *et al.*, "Large-scale topological and dynamical properties of the internet," *Phys. Rev. E*, vol. 65, no. 6, p. 066130, Jun 2002.

[24] V. Grolmusz, "A note on the pagerank of undirected graphs," May 2012. [Online]. Available: http://arxiv.org/abs/1205.1960

[25] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: A coauthorship network analysis," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 10, pp. 2107–2118, Oct. 2009.

[26] I. Swami *et al.*, "Attack robustness and centrality of complex networks," *PLoS ONE*, vol. 8, no. 4, April 2013.

[27] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks," *Phys. Rev. E*, vol. 65, no. 5, May 2002.

[28] A. Nucci, A. Sridharan, and N. Taft, "The problem of synthetically generating IP traffic matrices: Initial recommendations," *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 3, pp. 19–32, Jul. 2005.

[29] S. Even *et al.*, "On the Complexity of Timetable and Multicommodity Flow Problems," *SIAM Journal on Computing*, vol. 5, no. 4, 1976.

[30] J. Edmonds and R. M. Karp, "Theoretical improvements in algorithmic efficiency for network flow problems," *J. ACM*, vol. 19, no. 2, 1972.

[31] I. Mishkovski *et al.*, "Vulnerability assessment of complex networks based on optimal flow measurements under intentional node and edge attacks," in *Proc. of ICT Innovations 2009*. Springer, pp. 167–176.

[32] S. P. Borgatti, "Centrality and network flow," *Social Networks*, no. 27, pp. 55–71, 2005.

[33] R. Albert, H. Jeong, and A.-L. Barabasi, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, Jul. 2000.

[34] S. Trajanovski *et al.*, "Robustness envelopes of networks," *Journal of Complex Networks*, vol. 1, no. 1, pp. 44–62, 2013.