Fast Robustness Estimation in Large Social Graphs: Communities and Anomaly Detection

Fragkiskos D. Malliaros[1], Vasileios Megalooikonomou[1,2], and Christos Faloutsos[3]

[1]University of Patras, Greece
[2]Temple University, PA, USA
[3]Carnegie Mellon University, PA, USA

Introduction

- Given a large social graph, like a scientific collaboration network, what can we say about its robustness?
- Can we estimate a robustness index for a graph quickly?
- If the graph evolves over time, how these properties change?
- **Robust Graph:** Capable to retain its structure and its connectivity properties after the loss of a portion of its nodes and edges
- The property of robustness in real-world graphs is closely related to the notion of **community structure**
- We tackle the problem of estimating the robustness of a graph quickly, studying the **expansion properties**
- **Contributions:**
  - Fast robustness index
  - Patterns of real static and time-evolving social graphs
  - Anomaly detection

Preliminaries: Expansion Properties

- Good expander: Simultaneously sparse and highly connected
- Given a graph $G = (V, E)$, the *expansion* of any subset of nodes $S \subset V$, with size at most $\frac{|V|}{2}$, is defined as $\frac{|N(S)|}{|S|}$
- A graph is considered to have good expansion properties if every subset of nodes has good expansion (i.e., many neighbors)

Expansion, Robustness and Community Structure

- Why the expansion properties of a graph are important?
  - They offer crucial insights about the structure of a graph
  - They can act as a natural measure of the graph's robustness
  - Information about the presence or not of edges which can operate as bottlenecks inside the network
- Good expansion properties $\rightarrow$ high robustness, while poor expansibility reflects exactly the opposite behavior
- Connections with the community structure: good expansibility requires cuts with large size (i.e., large number of edges crossing the cut) $\rightarrow$ poor community structure
- The expansion properties of a graph can be approximated by the *spectral gap* $\Delta\lambda = \lambda_1 - \lambda_2$ of the adjacency matrix A
- Large $\Delta\lambda$ implies high robustness
  - **However**, it is not clear how large the spectral gap should be

Spectral Gap + Subgraph Centrality

- Combine the spectral gap with the subgraph centrality **[Estrada, Eur. Phys. J. B '06]**
- **Subgraph Centrality:** # of closed walks that a node participates

$$SC(i) = \sum_{j=1}^{|V|} u_{ij}^2 \sinh(\lambda_j), \ \forall i \in V$$

- Good expansion properties $\rightarrow$ High robustness $\rightarrow \lambda_1 \gg \lambda_2 \rightarrow$ $u_{i1}^2 \sinh(\lambda_1) \gg \sum_{j=2}^{|V|} u_{ij}^2 \sinh(\lambda_j)$
  - $SC(i) \approx u_{i1}^2 \sinh(\lambda_1), \ \forall i \in V \rightsquigarrow u_{i1} \propto \sinh^{-1/2}(\lambda_1)\, SC(i)^{1/2}$
  - Deviation from this behavior $\rightarrow$ existence (or lack thereof) of high robustness properties
- **Shortcoming:**
  - (i) Scalability issues (it requires all the pairs $(\lambda_i, u_i), \forall i \in V$)
  - (ii) It cannot be applied directly to bipartite graphs

Proposed Metric: Generalized Robustness Index r_k

Q: Can we efficiently approximate the *SC* of every node in the graph?

1. The eigenvalues of A follow a power-law distribution **[Faloutsos et al., SIGCOMM '99]**
2. The eigenvalues are almost symmetric around zero (except from the first few) **[Tsourakakis, ICDM '08]**
3. $\sinh(\cdot)$: odd function (i.e., $\sinh(-x) = -\sinh(x)$)

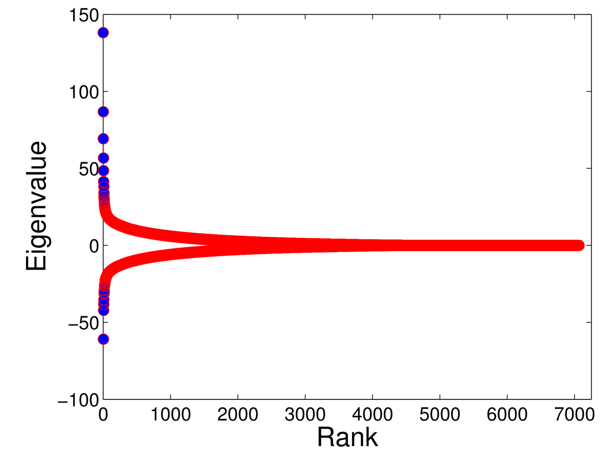


Figure: Skewed spectrum (Wiki-Vote)

$\rightsquigarrow$ Approximate the *SC* using **only the first top** k eigenvalues and their corresponding eigenvectors

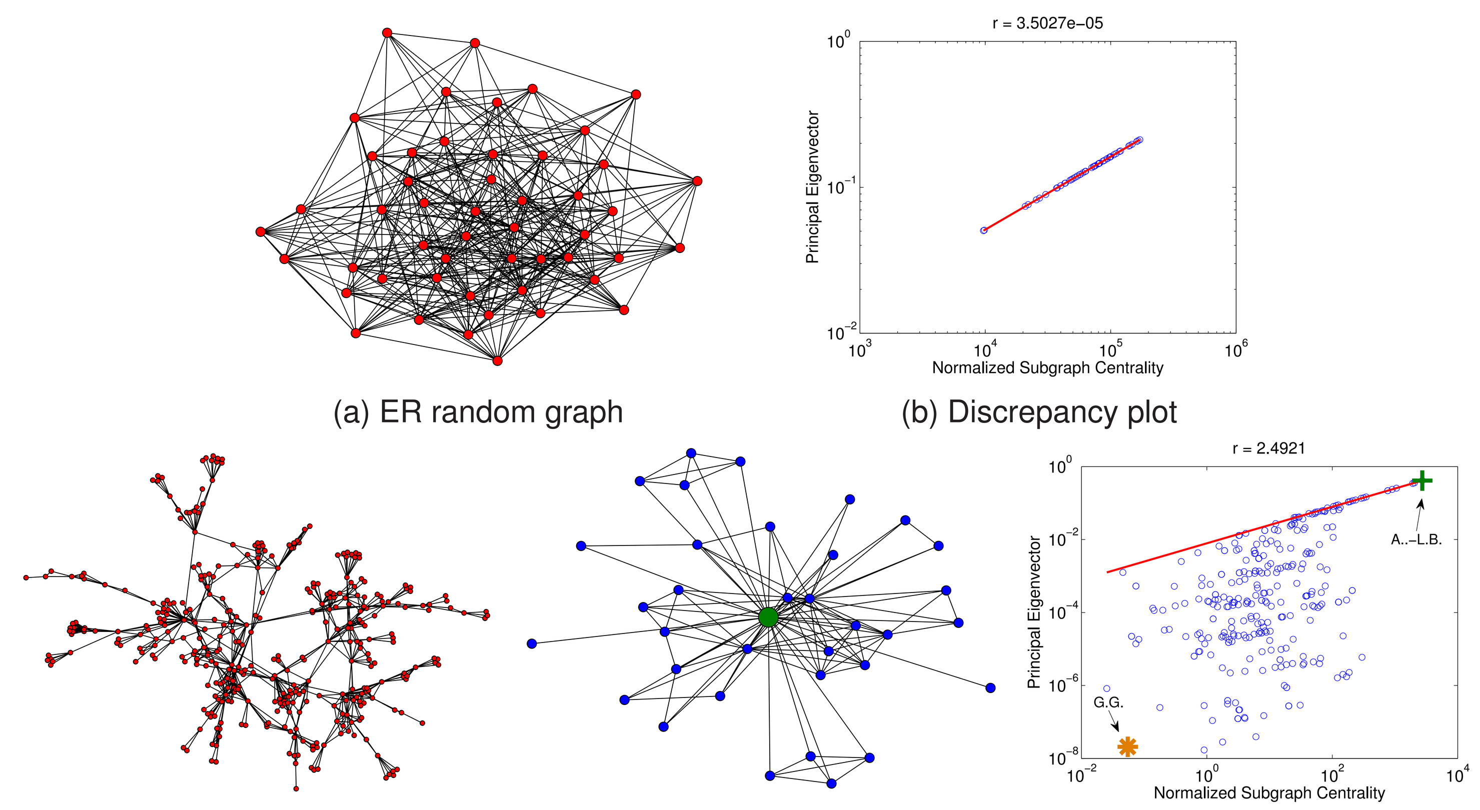
$$NSC_k(i) = \sum_{j=1}^{k} u_{ij}^2 \sinh(\lambda_j), \ \forall i \in V$$

- Generally, $k \ll |V|$ for real-world graphs

$$\rightsquigarrow r_k = \left(\frac{1}{|V|} \sum_{i=1}^{|V|} \left\{ \log(u_{i1}) - \left(\log(\sinh^{-1/2}(\lambda_1)) + \frac{1}{2}\log(NSC_k(i)) \right) \right\}^2 \right)^{1/2}$$

- Summarizes the robustness of a graph in a single number (smaller $r_k \rightarrow$ better robustness)

An Illustrative Example: Random vs. Real Graphs



(a) ER random graph (b) Discrepancy plot

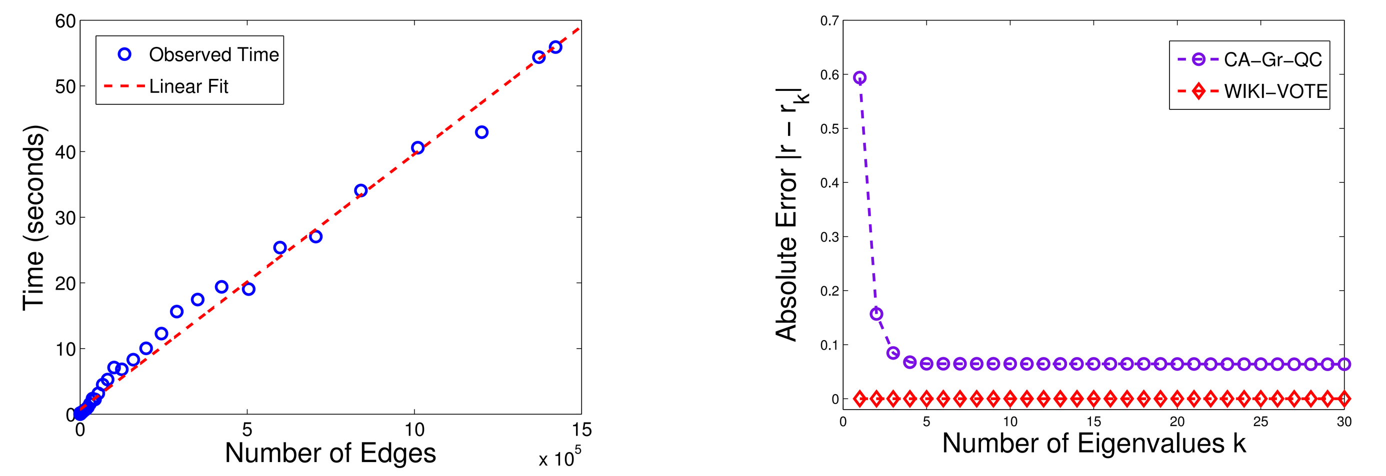
(c) Network science graph (d) A.-L. Barabasi's egonet (e) Discrepancy plot

Datasets

Graph	# Nodes	# Edges	Graph	# Nodes	# Edges
Epinions	75,877	405,739	CA-Astro-Ph	17,903	197,031
Email-EuAll	224,832	340,795	CA-Gr-Qc	4,158	13,428
Slashdot	77,360	546,487	CA-Hep-Th	8,638	24,827
Wiki-Vote	7,066	100,736	Dblp	404,892	1,422,263
Facebook	63,392	816,886	Cit-Hep-Th	26,084	334,091
YouTube	1,134,890	2,987,624			

Effectiveness and Scalability of r_k index

Q: How effective and scalable (efficient) is the proposed r_k index?

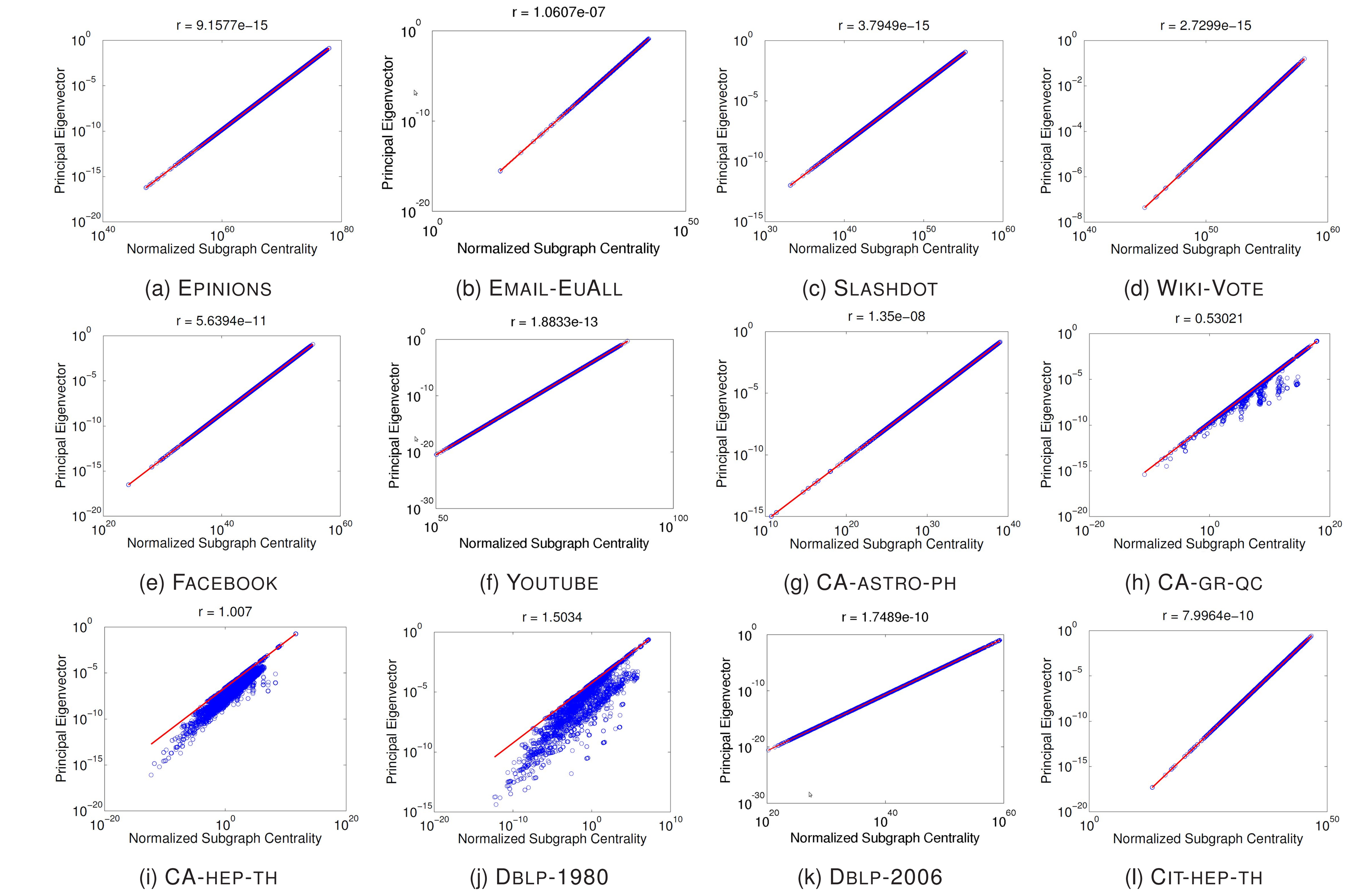


(a) Scalability (Dblp – $k = 30$ eigs) (b) Effectiveness (CA-Gr-Qc & Wiki-Vote)

- The r_k index scales linearly with respect to the number of edges
- Only a few eigenpairs are enough to achieve a very good approximation of the robustness index

Robustness of Large Static Graphs: High Robustness Pattern

Q: What can we say about the robustness of large social graphs?

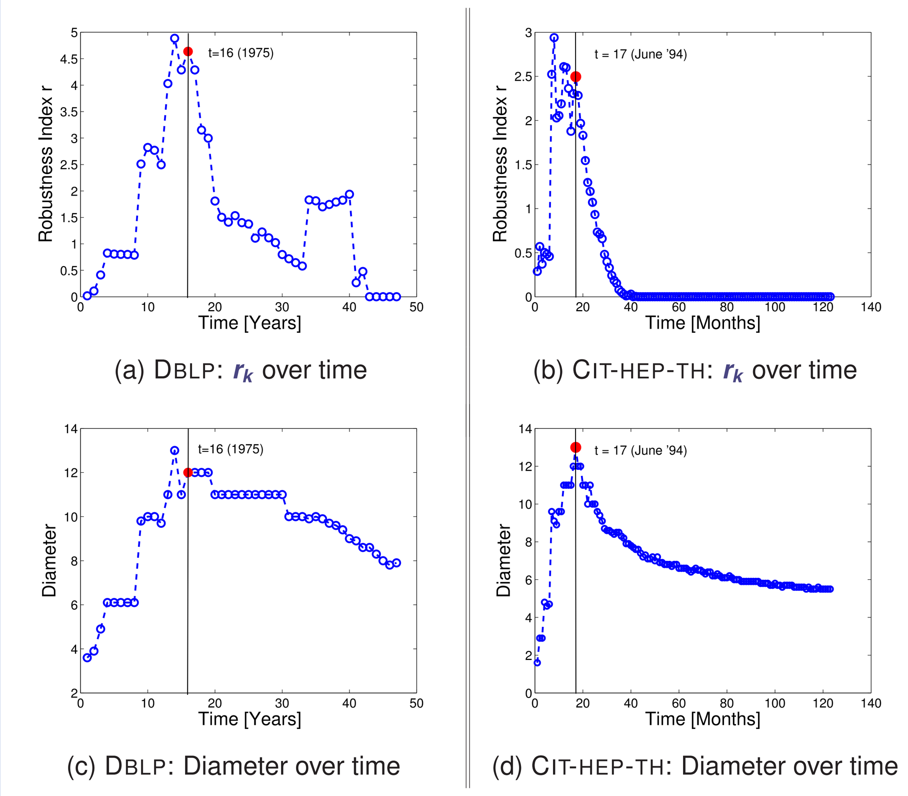


(a) Epinions (b) Email-EuAll (c) Slashdot (d) Wiki-Vote
(e) Facebook (f) YouTube (g) CA-Astro-Ph (h) CA-Gr-Qc
(i) CA-Hep-Th (j) Dblp-1980 (k) Dblp-2006 (l) Cit-Hep-Th

Figure: **Discrepancy plots**: Principal eigenvector vs. *NSC* in log-log scales. Almost all graphs tend to be extremely robust (linearity, r_k close to zero)

Large Social Graphs: Good expansion properties $\rightarrow$ High robustness $\rightarrow$ Not a clear modular structure

Time Evolving Graphs: Fragility Evolution Pattern



(a) Dblp: r_k over time (b) Cit-Hep-Th: r_k over time

(c) Dblp: Diameter over time (d) Cit-Hep-Th: Diameter over time

Figure: Fragility evolution pattern: The spike of the r_k index aligns with the diameter's spike

Q: How the robustness index r_k of a graph changes over time?

- Study the *fragility evolution* of a graph

General Observation:

- At the first time points $r_k \nearrow$ gradually $\rightarrow$ Low robustness $\rightarrow$ Good community structure
- After a specific time point, r_k starts $\searrow$ gradually $\rightarrow$ The graphs tend to become more robust
- The time point that r_k changes, corresponds to the **gelling poing [McGlohon et al., KDD '08]**

The fragility evolution pattern can be considered as a natural explanation for the structural differences (regarding robustness and community structure) between different scale graphs

Anomaly Detection

Q: Can we spot anomalies over time using the r_k index?

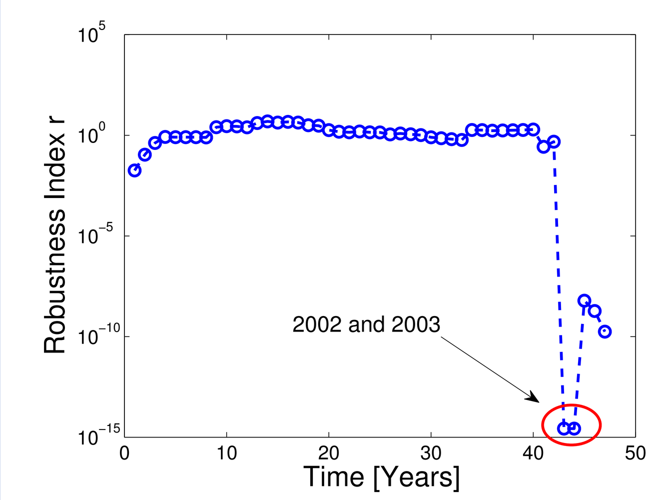


Figure: Fragility evolution of the Dblp graph (lin-log scales)

- Examine the r_k index over time, trying to identify and track abrupt changes and deviations
- Sudden deviations from the fragility evolution pattern can possibly correspond to anomalies
- Dblp graph: Strange behavior of the r_k index for 2002 and 2003
- These two time graphs are outliers
  - After 2001 a large number of new publications were introduced $\rightarrow$ Robustness $\nearrow \rightarrow r_k \searrow$
  - After 2002-2003 new research fields are covered from Dblp $\rightarrow$ New fields formed new communities $\rightarrow r_k \nearrow$

Acknowledgements

Data: SNAP, MPI-SWS, NEU, DBLP

Funding: NSF (Grant No. IIS1017415), ARL (No. W911NF-09-2-0053)