Anomaly Detection and Localization in Graphs

David J. Marchette

Naval Surface Warfare Center
This work was supported by the NSWCDD ILIR Program

2011 HUIC Mathematics and Engineering Conference

Outline

Collaborators

- Carey Priebe, JHU.
- Glen Coppersmith, JHU.
- Andrey Rukhin, NSWC.
- Many others.

Acknowledgment: This work funded in part by the Office of Naval Research In-House Laboratory Independent Research program and by the Naval Innovative Science and Engineering Program (Section 219).

Outline

Communications Graphs

- The problem we are investigating comes from the analysis of social networks, in particular communications graphs.
- Consider a graph in which the vertices are people and there is an edge between two people if they communicate (phone, email, personal contact, et cetera).
- The basic assumption (which we'll make precise later) is that the graph is *homogeneous* except for a small group of exceptional individuals.
  - The exceptional group communicate amongst themselves more than is typical.
  - This is measured as the number of edges in the induced subgroup is larger than is typical.
- We will set up a model in which we can make these precise, and give methods and results to detect and localize the exceptional group.

Definitions

- All graphs $G = (V, E)$ are undirected and simple.
- $|V| = n$, $|E| = s$. The degree of a vertex v will be denoted d_v.
- Define the closed neighborhood of v as $N[v] = \{w \in V | vw \in E\} \cup \{v\}$.
- The induced subgraph of a set $U \subset V$ is denoted $\Omega(U)$ and consists (U, F), where F contains all edges in E between elements of U.

Random Models

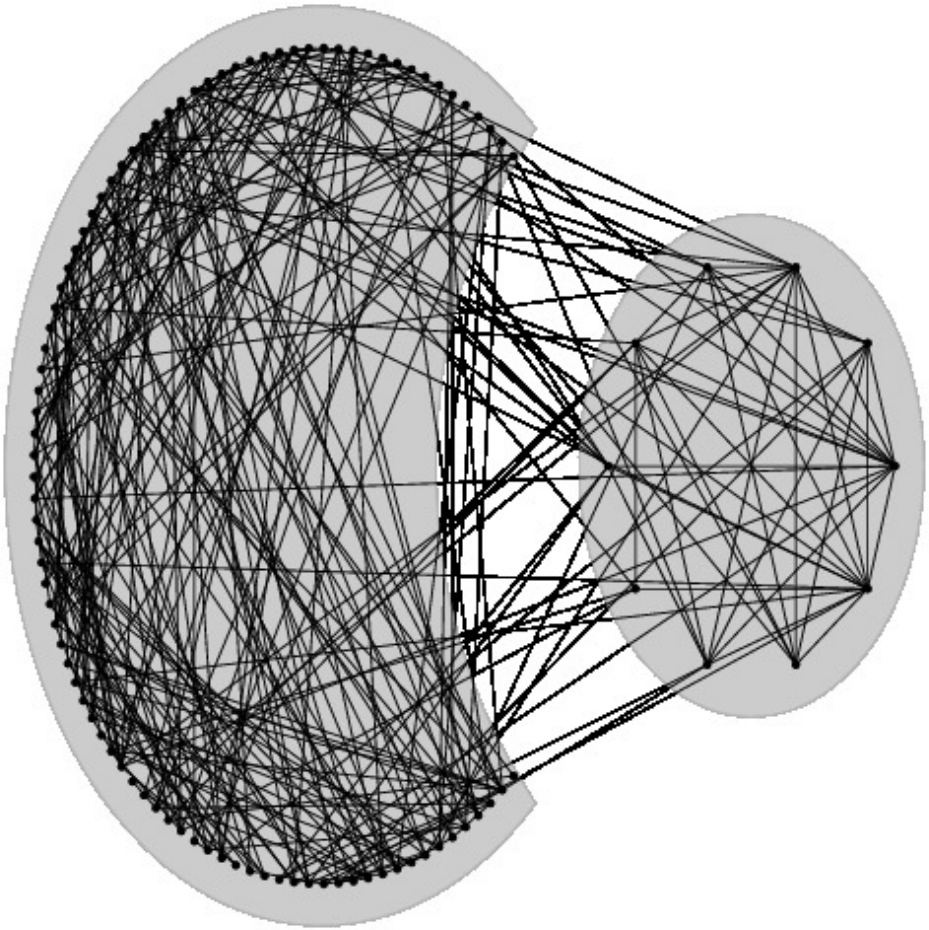
- The null model will be the Erdös-Renyi random graph:

$$P[uv \in E] = p,$$

  with all edges independent. We'll write $G \sim \text{ER}$.
- The alternative will be $\kappa(n, p, m, q)$: Let $V = V_1 \cup V_2$ with $|V_1| = n - m$, $|V_2| = m$.

$$P[uv \in E] = \begin{cases} p & u \in V_1 \text{ or } v \in V_1 \\ q & u, v \in V_2 \end{cases}$$

A κ Graph



The Random Dot Product Graph (RDPG)

- We will make use of a latent position model in the localization work.
- $G = (V, E; X)$ where $X \in \amalg^n \Delta^d$ (or $D^d \cap \mathbb{R}^{+d}$).
- We assume the edge probabilities are given by:

$$P[uv \in E] = X_u^T X_v.$$

- We will estimate X, then use this to try to localize the anomalous vertices.
- Note: we assume the rows of X correspond to the vertices.

Estimating the Latent Attributes

- We can estimate the latent vectors X by making the observation that if we could augment the adjacency matrix A of G with $X_u^T X_u$ (call this $\widetilde{A}$) the spectral theorem provides the optimal solution X to minimizing

$$\|\widetilde{A} - XX^T\|.$$

- It turns out, a good estimate of $X_u^T X_u$ is

$$\frac{d_u}{n-1},$$

where d_u is the degree of u.

Outline

Local vs Global

- A global statistic is one that uses the entire graph:
  - Size – number of edges.
  - Number of triangles (copies of K_3)
  - Clustering coefficient.
  - Average path length.
- A local statistic focuses on local regions:
  - Maximum degree.
  - Scan statistic (defined below).
- Although these are computed across the full graph, they use only local information.

Scan Statistic

- A locality statistic is a graph invariant computed on the induced subgraph of the closed neighborhood (or any other definition of "local" in the graph) of each vertex:

$$T(v) = T(\Omega(N[v])).$$

- A scan statistic is the maximum locality statistic:

$$\widetilde{T} = \max_{v} T(v).$$

- We consider the size (number of edges) of the subgraph.
- Note that by keeping track of which v is associated with $\widetilde{T}$ we can use this for localization.

Outline

Detecting an Anomaly

- We perform the hypothesis test:

$$\begin{aligned} H_0 &: G \sim \text{ER} \\ H_1 &: G \sim \kappa \end{aligned}$$

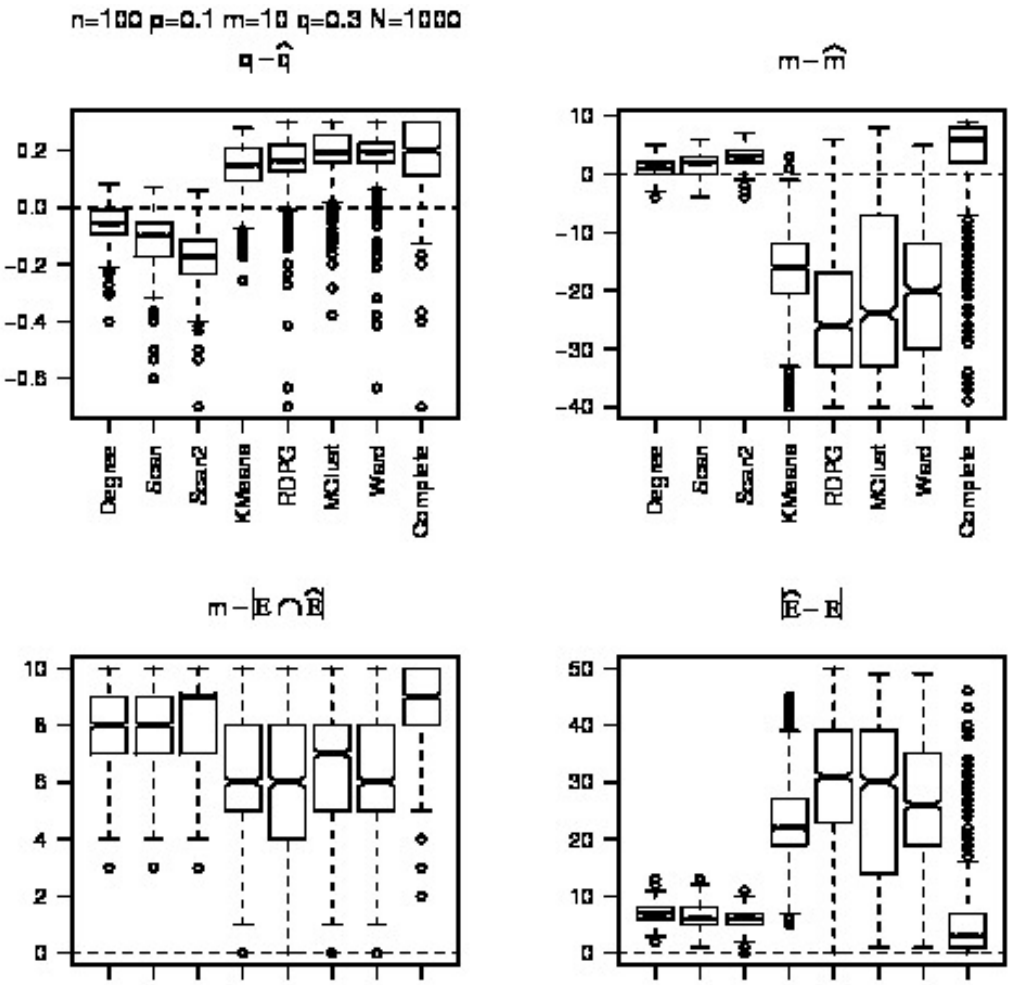
- Rukhin showed that (asymptotically, $m = O(\sqrt{n})$):
  - The number of triangles in the graph has more power than the size of the graph.
  - Size has more power than max degree, but for $n < 10^{24}$ max degree has more power.
  - The scan statistic has more power than max degree.
- Simulations show that scan statistic is generally optimal, among these statistics, for detecting this type of anomaly.
- This is as expected.

Outline

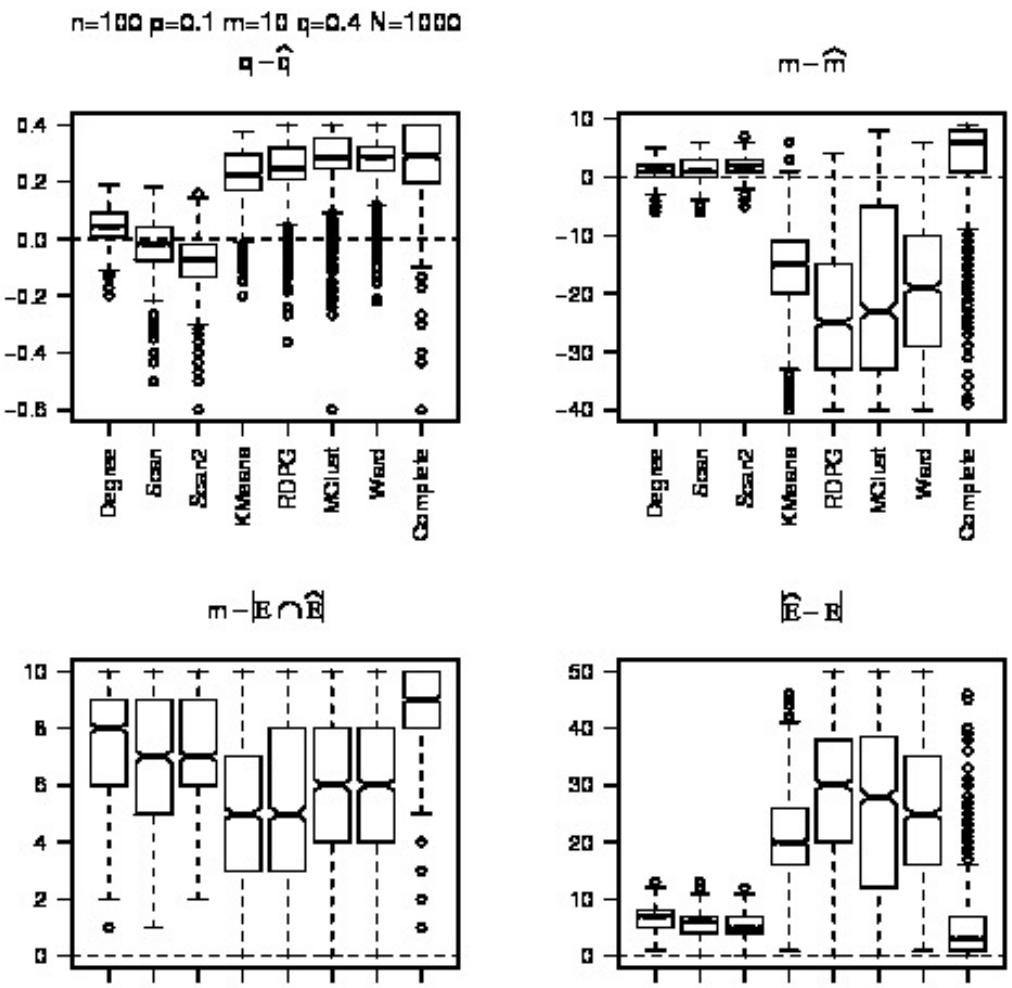
Localizing the Detection

- Assume we have detected the existence of an anomaly.
- We can estimate the position of the anomaly in various ways. We will consider:
  - Methods that use the detection statistics for localization:
    - vertex with maximum degree.
    - scan region of maximum size.
  - Methods that focus exclusively on the graph:
    - Cluster the vertices in the graph using various methods.
    - Cluster the RDPG embedding.
- We'll look at the error in estimating m and in selecting the correct m vertices.

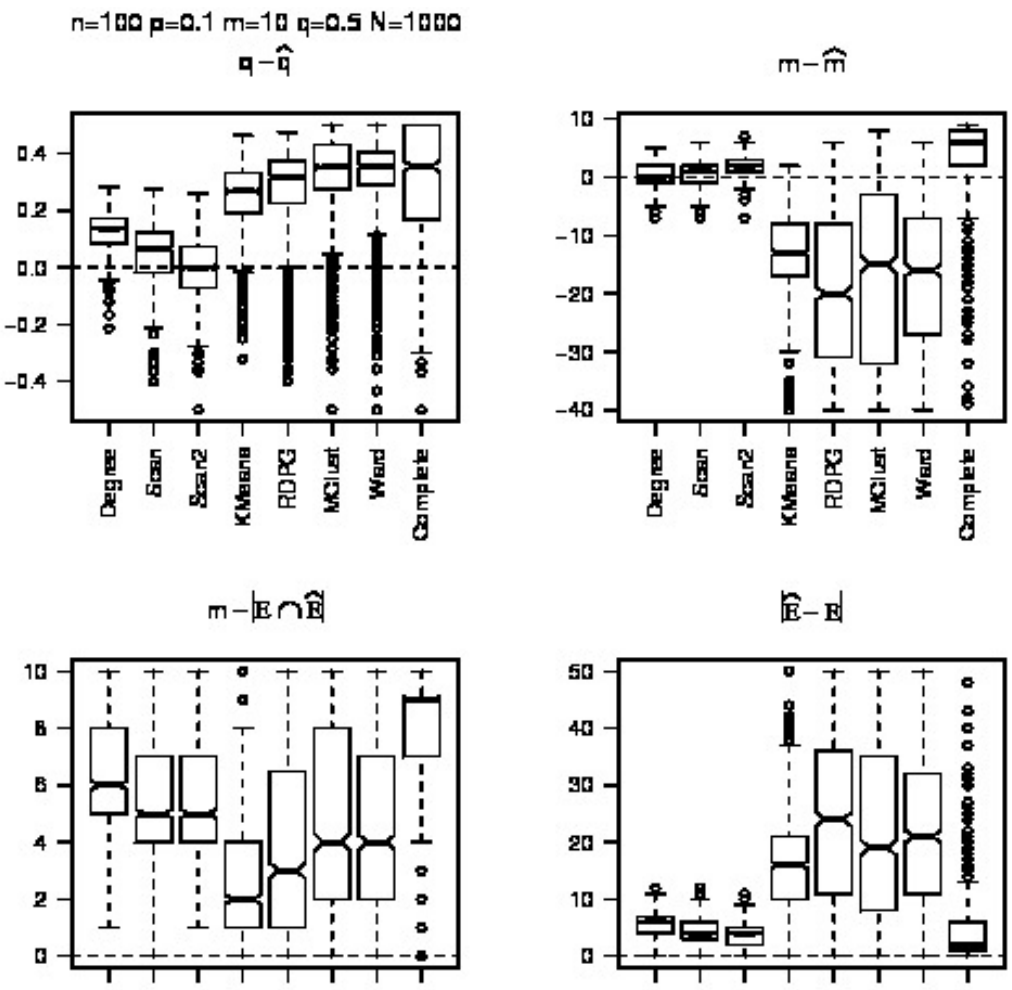
Examples


n=100 p=0.1 m=10 q=0.3 N=1000
q−q̂
m−m̂
m−|E∩Ê|
Ê−E
Degree
Scan
Scan2
KMeans
RDPG
MClust
Ward
Complete

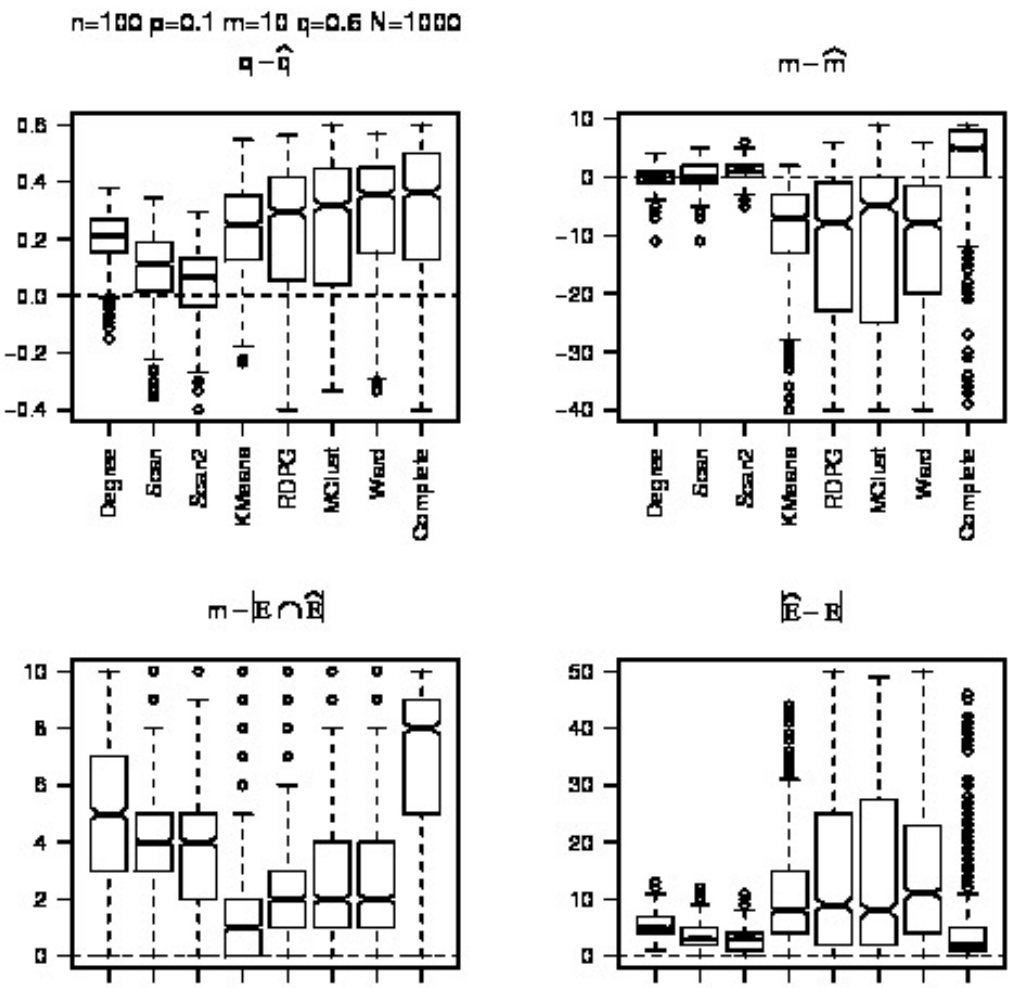
Examples


n=100 p=0.1 m=10 q=0.4 N=1000
q−q̂
m−m̂
m−|E∩Ê|
Ê−E
Degree
Scan
Scan2
KMeans
RDPG
MClust
Ward
Complete

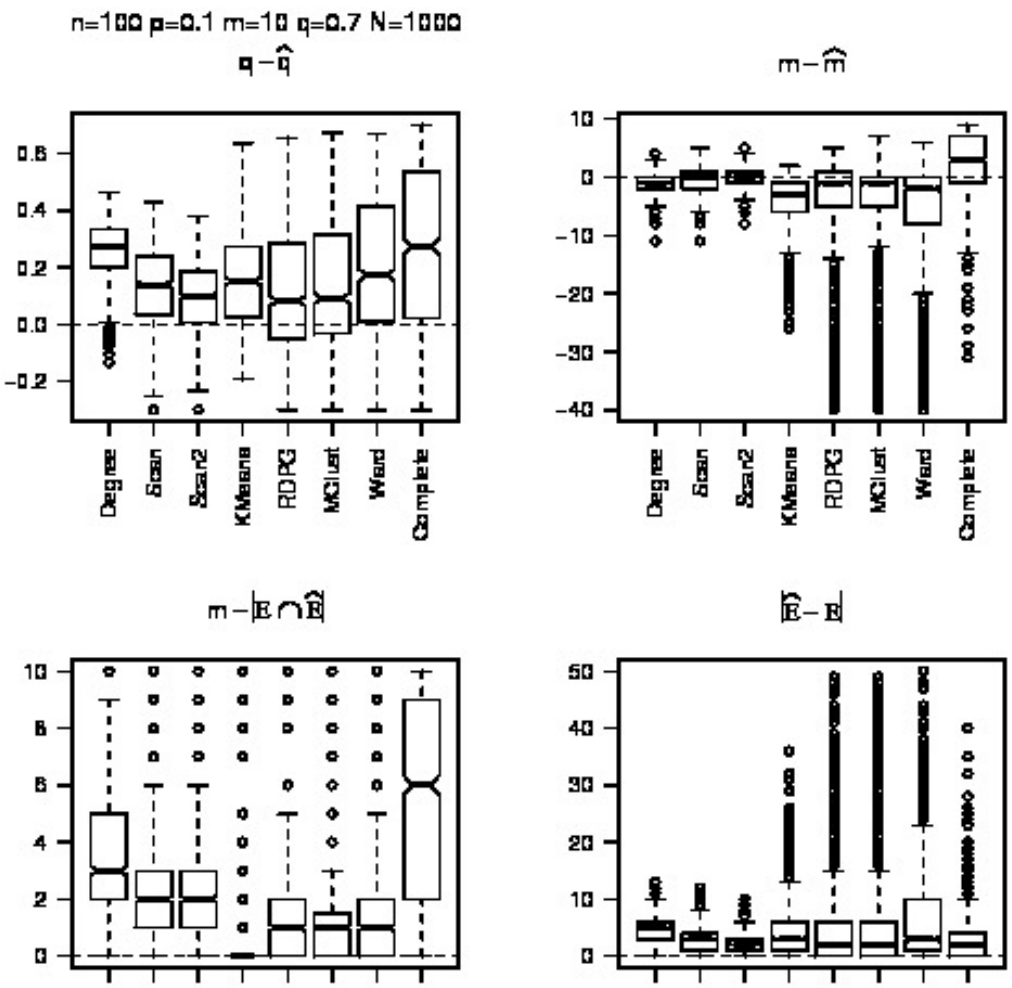
Examples


n=100 p=0.1 m=10 q=0.5 N=1000
q − q̂
m − m̂
m − |E ∩ Ê|
|Ê − E|
Degree
Scan
Scan2
KMeans
RDPG
MClust
Ward
Complete

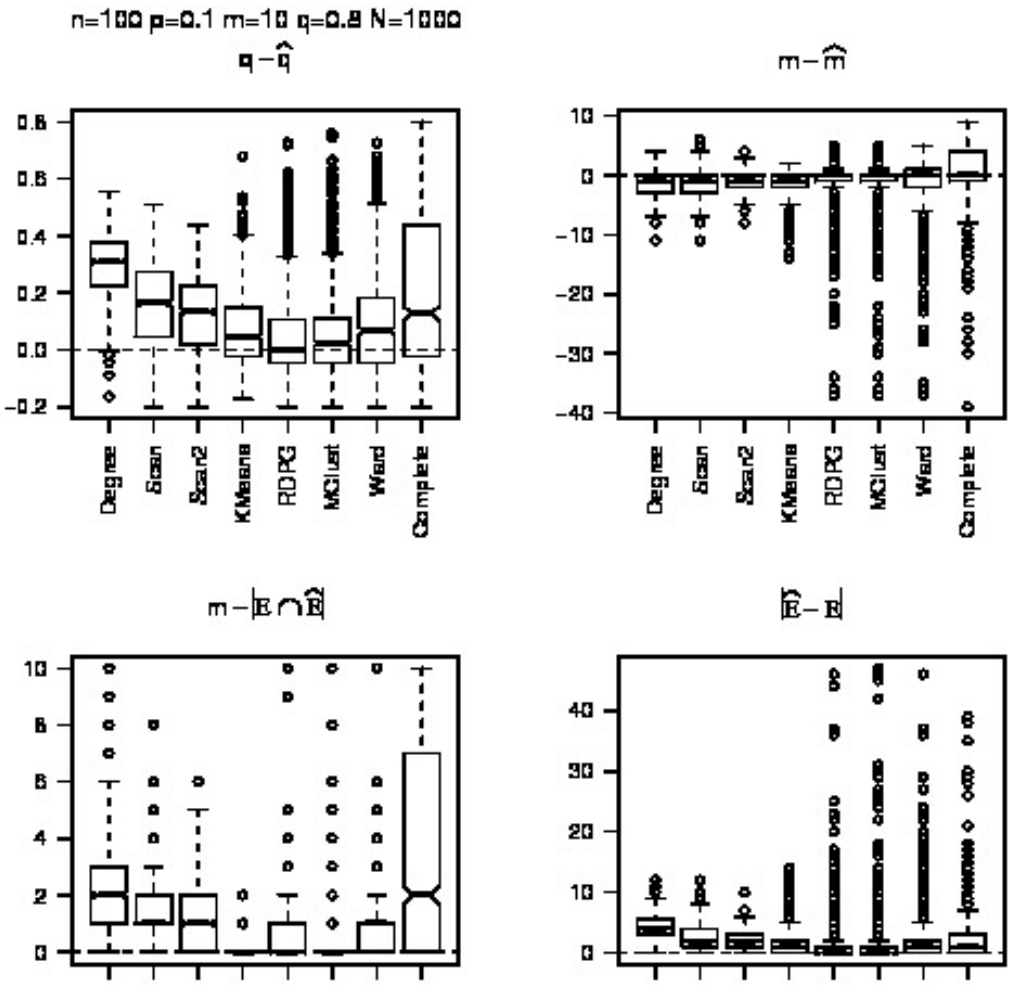
Examples


n=100 p=0.1 m=10 q=0.6 N=1000
q−q̂
m−m̂
m−|E∩Ê|
|Ê−E|
Degree
Sscan
Scan2
KMeans
RDPG
MClust
Ward
Complete

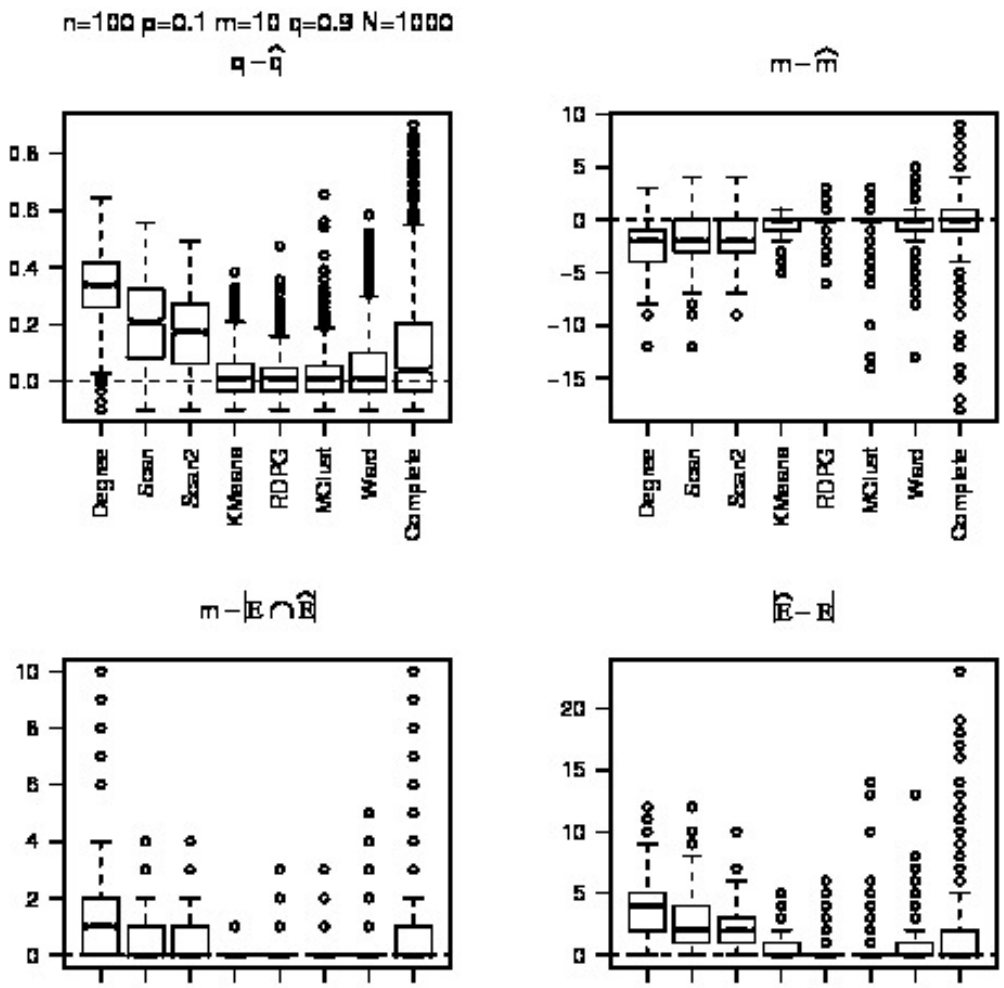
Examples


n=100 p=0.1 m=10 q=0.7 N=1000
q−q̂
m−m̂
m−|E∩Ê|
|Ê−E|
Degree
Scan
Scan2
KMeans
RDPG
MClust
Ward
Complete

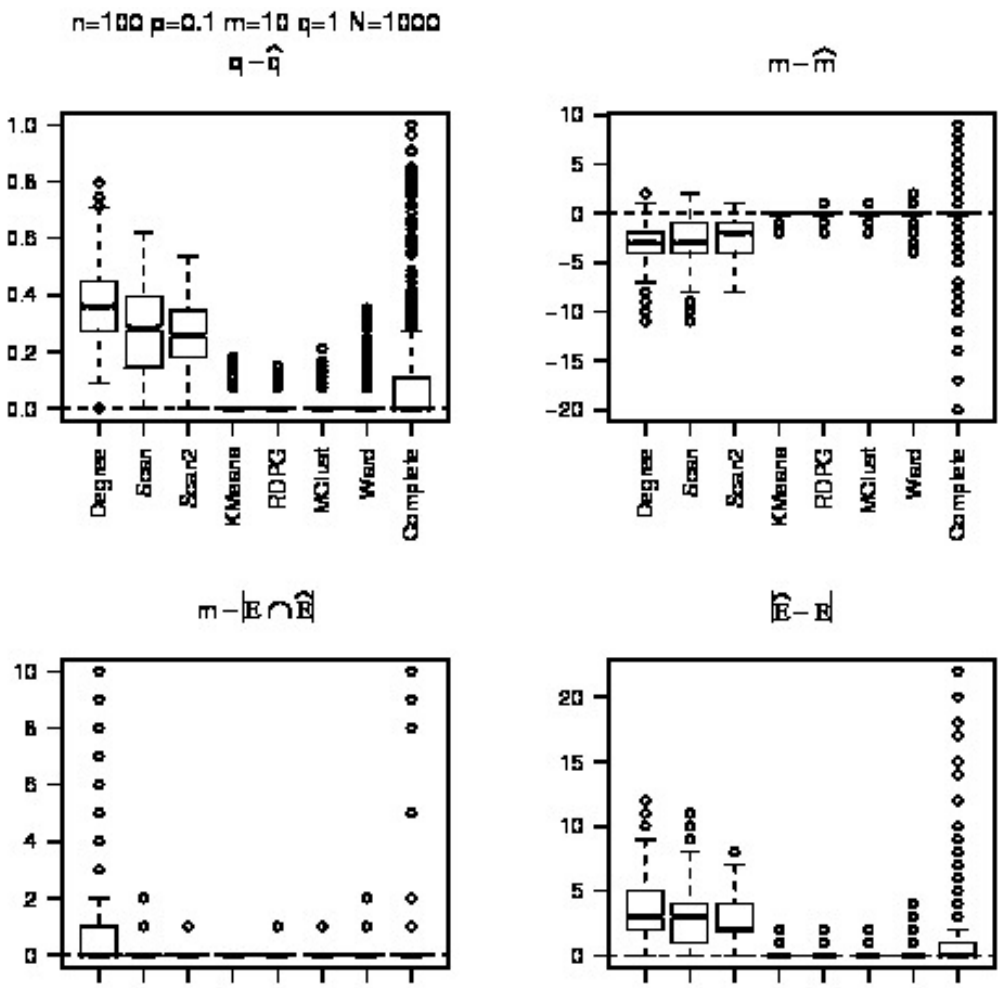
Examples


n=100 p=0.1 m=10 q=0.9 N=1000
q−q̂
m−m̂
m−|E∩Ê|
Ê−E
Degree
Sscan
Scan2
KMeans
RDPG
MClust
Ward
Complete

Examples


n=100 p=0.1 m=10 q=0.9 N=1000
q−q̂
m−m̂
m−|E∩Ê|
|Ê−E|
Degree
Sscan
Scan2
KMeans
RDPG
MClust
Ward
Complete

Examples


n=100 p=0.1 m=10 q=1 N=1000
q−q̂
m−m̂
m−|E∩Ê|
|Ê−E|
Degree
Scan
Scan2
KMeans
RDPG
MClust
Ward
Complete

Outline

Discussion

- The detection and localization of anomalies in graphs is important for many important applications.
- We have shown that scan statistics and the random dot product graph are powerful tools for performing these detections and localizations.
- The random dot product model can be extended to large graphs through sparse matrix operations, provided the graphs are sufficiently sparse, which is often the case for real-world communications graphs and social networks.

Future Work

- We are extending these results to attributed graphs.
  - Edges have an attribute such as the topic of communication.
  - Vertices have an attribute such as group membership or measurements of interests or other attributes associated with social position.
- Extensions to multi- and hyper-graphs are possible and preliminary work has been done on these.
- Work on the detection and localization of multiple anomalies is also ongoing.