Statistical Models and Methods for Anomaly Detection in Large Graphs

Nicholas Arcolano and Benjamin A. Miller
MIT Lincoln Laboratory

2012 SIAM Annual Meeting

9–13 July 2012

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
244 Wood Street, Lexington, MA 02420-9108

This work is sponsored by the Intelligence Advanced Research Projects Activity (IARPA) under Air Force Contract FA8721-05-C-0002. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Outline

- **An anomaly detection framework for massive graphs**

- **Modeling attributed graphs using generalized linear models**

- **Empirical results**

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Introduction

- **Graphs and networks constitute a valuable theoretical framework for modeling and analyzing relational data**

- **“Very large” or “massive” graphs:**
  - **Arise from “very large” data sets**
  - **Nodes can number in the millions to billions (e.g. document and media databases, social networks, the Internet) or even larger (e.g. biological and molecular interaction networks)**

- **We would like to use these data to perform classical types of analysis, i.e. signal processing**
  - **Hypothesis testing, parameter estimation, classification, time series analysis, anomaly and change detection**

- **However, there are significant challenges:**
  - **Graphs are inherently combinatorial, non-Euclidean**
  - **Extensions of traditional theory to graphs is lacking or cumbersome**
  - **Scale of massive graphs imposes substantial constraints on computation**

Anomaly Detection

- **Classical anomaly detection**
  - **Observed data vectors** $\mathbf{y}_1, \ldots, \mathbf{y}_n$
  - **We assume that most data are drawn from an unknown joint distribution** $p(\mathbf{y}_1, \ldots, \mathbf{y}_n)$**, although some points may not be**

- **Want to determine which (if any) points deviate from the model**
  - **Detection: do any anomalous points exist?**
  - **Classification: if so, which ones are they?**

- **A common approach: residuals analysis**
  - **Posit a model** $p(\mathbf{y}_1, \ldots, \mathbf{y}_n)$
  - **Estimate model** $\hat{p}(\mathbf{y}_1, \ldots, \mathbf{y}_n)$
  - **Compute expected observations under the model** $\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_n$
  - **Detect anomalies based on residual errors, e.g.**

$$\epsilon_i = \|\mathbf{y}_i - \bar{\mathbf{y}}_i\|_2 \qquad T(\mathbf{y}_1, \ldots, \mathbf{y}_n) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}_i\|_2^2$$

An Anomaly Detection Framework for Massive Graphs

- **We wish to extend this classical framework to massive graphs**
  - **Given an observed graph G with n nodes**
  - **Want to know if an anomalous subgraph exists within G (and if so, where is it?)**

- **Residuals-based anomaly detection**
  - **Observed adjacency matrix $\mathbf{A} = \{a_{ij}\}$**
  - **Estimate of expected adjacency matrix $\bar{\mathbf{A}}$**
  - **Analyze error matrix $\mathbf{E} = \mathbf{A} - \bar{\mathbf{A}}$ to detect and identify anomalous subgraphs**

- **Challenges**
  - **Need to be able to estimate model parameters efficiently**
  - **Detection: need to be able to compute test statistic (e.g. $\|\mathbf{E}\|$) efficiently**
  - **Classification: need to be able to identify subgraph with large residual error efficiently (e.g. using sparse spectral methods)**

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

The Chung-Lu Random Graph Model

- **Definition**
  - **Consider a simple random graph G with n nodes**
  - **For $i < j$, let the adjacency matrix $\mathrm{A} = \{a_{ij}\}$ of G be Bernoulli RVs with probability**

$$\Pr\left(a_{ij} = 1\right) = p_{ij} = w_i w_j$$

  - **Thus, we have $\mathbb{E}\left(\mathbf{A}\right) = \mathbf{P} = \mathbf{w}\mathbf{w}^T$**

- **Given an observed graph, a common estimator for w_i is**

$$\hat{w}_i = \frac{k_i}{\sqrt{\sum_{j=1}^{n} k_j}} = \frac{k_i}{2m},$$

where k_i is the i-th observed degree and m is the number of edges

- **Consequently, an estimate of $\mathbb{E}\left(\mathbf{A}\right)$ under the Chung-Lu model is**

$$\bar{\mathbf{A}} = \hat{\mathbf{w}}\hat{\mathbf{w}}^T = \frac{\mathbf{k}\mathbf{k}^T}{2m}$$

The Chung-Lu Model, Modularity, and Residuals

- **Applying the Chung-Lu model within the anomaly detection framework yields the residuals matrix**

$$\mathbf{E} = \mathbf{A} - \bar{\mathbf{A}} = \mathbf{A} - \frac{\mathbf{k}\mathbf{k}^T}{2m},$$

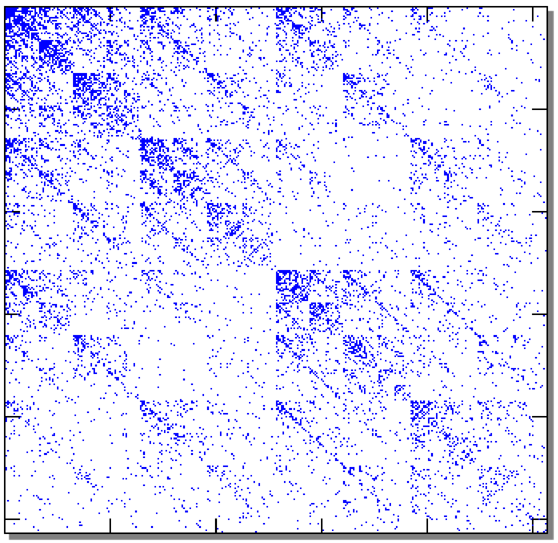
which is equivalent to the modularity matrix of G

- **Thus, when A is sparse**
  - **Residuals matrix is the sum of a sparse and a rank-1 matrix**
  - **We can compute norms, eigenvalues, and eigenvectors of E efficiently using sparse methods**

Special structure in the residuals matrix enables anomaly detection large graphs

Anomaly Detection Example

RESIDUALS CONSTRUCTION → **MATRIX DECOMPOSITION** → **FEATURE SELECTION** → **DETECTION** → **CLASSIFICATION**

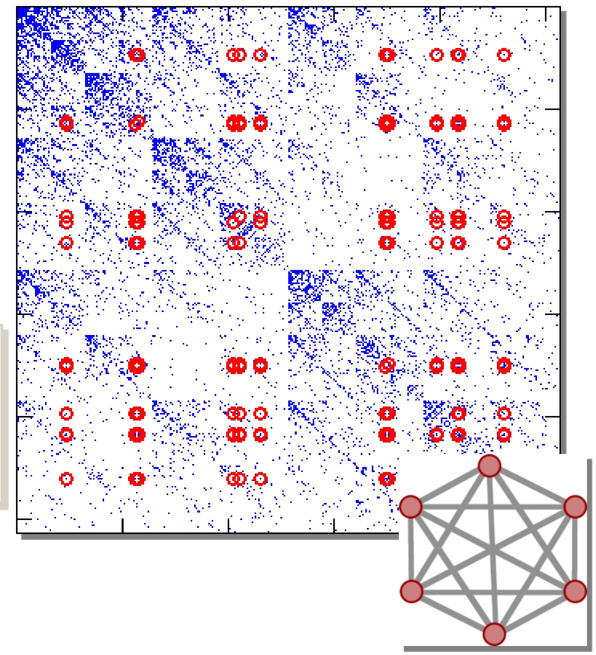


INPUT

$\mathbb{A}$, adjacency matrix representation of G

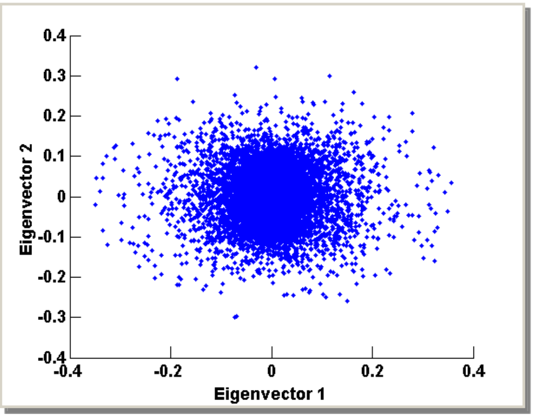
OUTPUT

***Set of vertices identified as* belonging to anomalous subgraph**



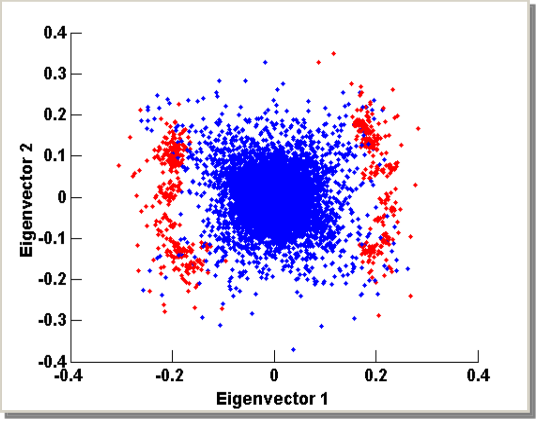
Anomaly Detection Example

H_0



Background (normal behavior) only

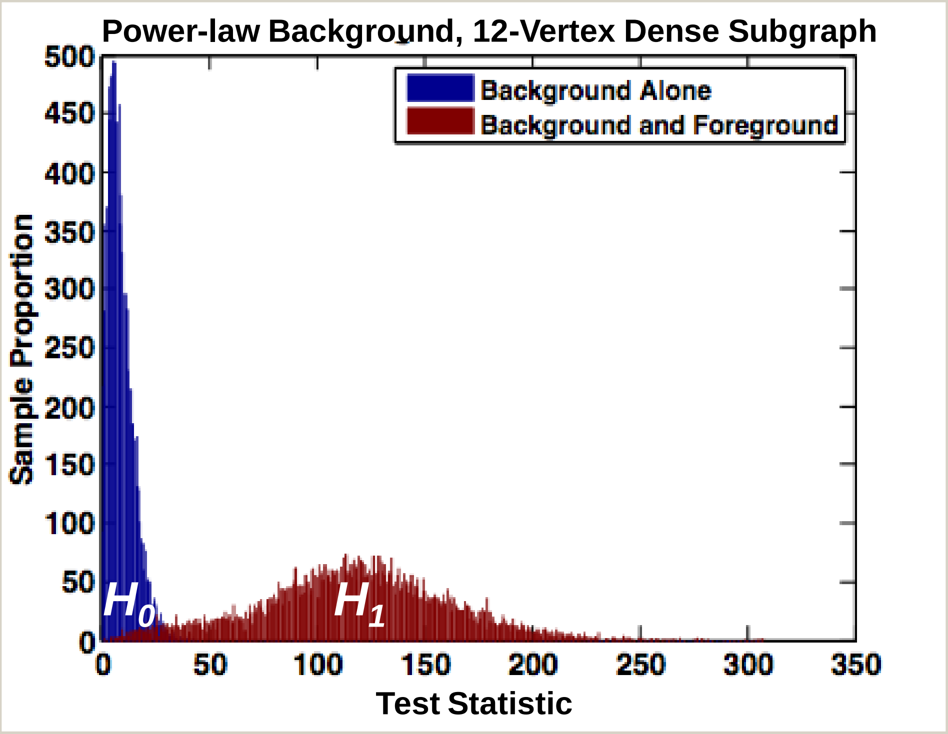
H_1



Background and foreground (includes anomalous behavior)

TEST STATISTIC:

SYMMETRY OF THE PROJECTION ONTO SELECTED COMPONENTS

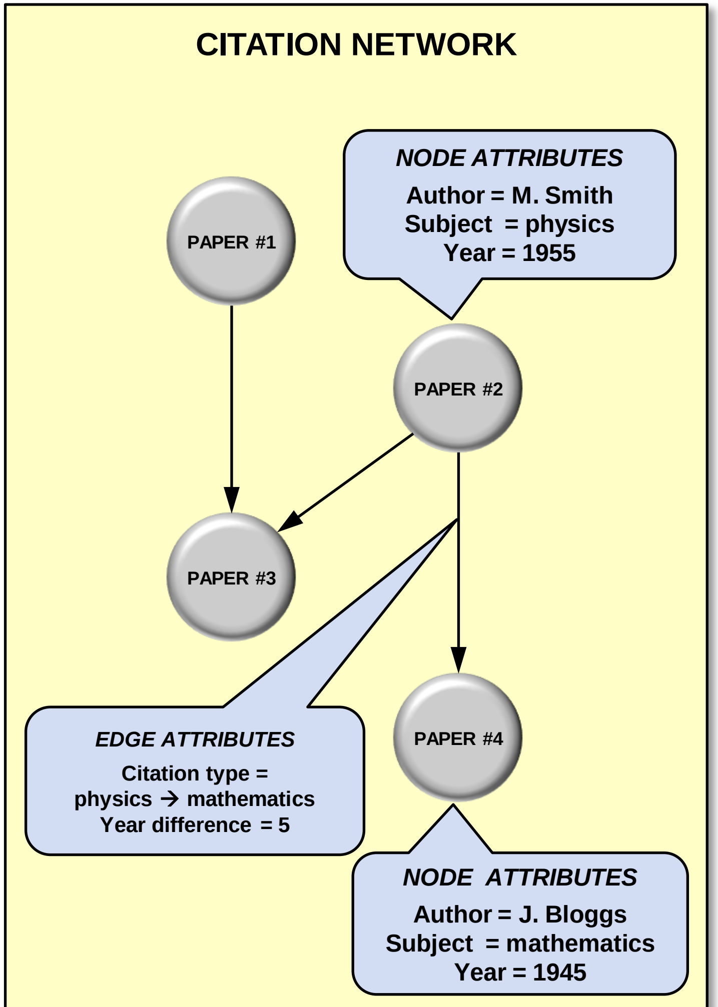


Distributions under H_0 and H_1 are well separated

Outline

- An anomaly detection framework for massive graphs
- Modeling attributed graphs using generalized linear models
- Empirical results

Attributed Graphs



- Typically, we define a graph as having only nodes and edges
- An attributed graph also has “attributes”, i.e. additional information about nodes and edges
  - In practice, we often construct a graph from raw data
  - Data not used to construct the graph can still be included as attributes
- We would like to perform anomaly detection using attributed graphs
  - Still need a model that admits special structure to enable computation

Generalized Linear Models

- **One approach to modeling attributed graphs is using generalized linear models (GLMs)**
  - **Used widely in classical statistics (e.g. logistic regression)**
  - **Increasingly used for modeling networks with attributes**

- **Definition**
  - **Let $\mathrm{X}_1, \ldots, \mathrm{X}_p$ denote matrices of covariates (attributes) for each potential edge**
  - **Conditioned on the covariates, assume the edges in G are generated by independent Bernoulli trials**
  - **We assume the expected value of the adjacency matrix is given by**

$$\mathbb{E}\left(a_{ij}\right) = p_{ij} = g\left(\sum_{k=1}^{p} \beta_k x_{ij}^{(k)}\right),$$

where $g : \mathbb{R} \to (0,1)$ is a link function such as the logistic function

$$g(t) = \frac{1}{1+\exp(-t)}$$

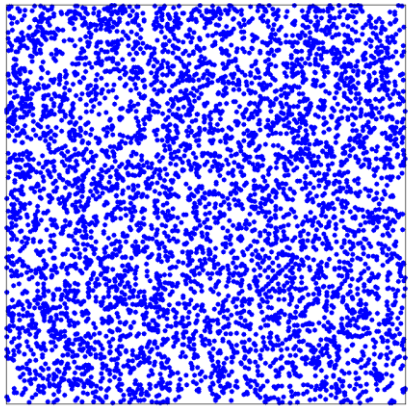
Generalized Linear Models for Networks: Advantages/Disadvantages

- **Advantages**
  - Allows us to incorporate covariates in the data to model attributed graphs
  - Extends a well-understood area of classical statistics
  - Model estimation is (somewhat) tractable
    - Maximum-likelihood estimate of GLM weights can be obtained via convex optimization
    - Gradient and Hessian of ML cost function can be expressed in closed form
    - Closed-form expressions for parameter estimates available in certain special cases

- **Disadvantages**
  - Estimation is more computationally demanding than simpler models
  - Exact estimation may not be possible for sufficiently large networks
  - Still need special structure to avoid producing a dense, high-rank estimate of expected adjacency matrix

Relationships Between GLM and Other Common Graph Models

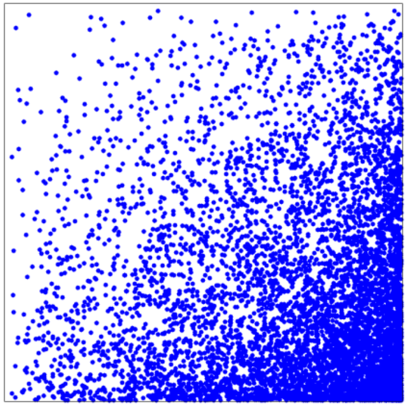
Erdős-Rényi



Edge probability constant across all pairs

$$E[a_{ij}] = \frac{1}{1+\exp(-\beta)}$$

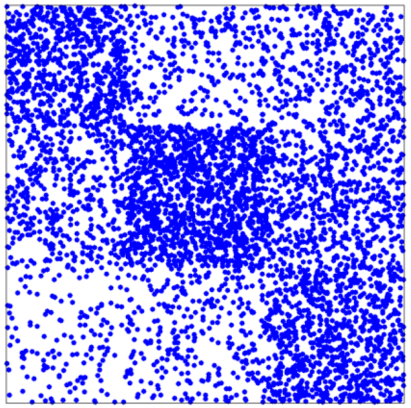
Chung–Lu *



Edge probability proportional to product of expected degrees

$$E[a_{ij}] = \exp\left((\delta_i + \delta_j)^T \beta\right)$$

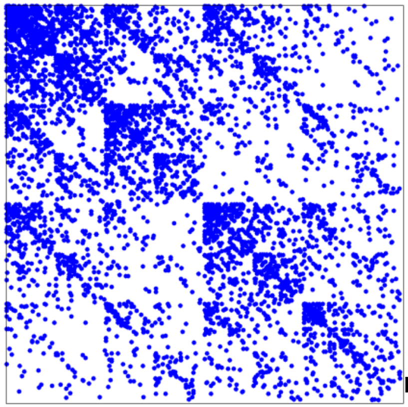
Stochastic Blockmodel*



Edge probabilities depend on groups of source and destination vertices

$$E[a_{ij}] = \exp\left(\delta_{c(i,j)}^T \beta\right)$$

Stochastic Kronecker Graph*



Probability matrix defined by the *n*-fold Kronecker product of a base matrix

$$E[a_{ij}] = \exp\left(f^T(i,j)\beta\right)$$

(*f* maps *i* and *j* to the number of times each section of the base matrix was used when generating probabilities recursively)

*These models do not restrict expected values to be in (0, 1), though the exponential approaches the logistic for small values

Generalized Linear Models for Networks: Example

- **Citation network**
  - **Database of n publications with associated bibliographic data (e.g. author, subject, journal)**
  - **Directed, unweighted graph with adjacency matrix $\mathbf{A} = \{a_{ij}\}$, where $a_{ij} = 1$ indicates document i cites document j**

- **Covariates**
  - **Let c be the number of different subjects**
  - **Each edge has $p = c^2$ categorical covariates indicating corresponding subject pair**

$$x_{ij}^{(k)} \in \{0, 1\} \qquad \sum_{k=1}^{p} x_{ij}^{(k)} = 1$$

- **Generalized linear model**
  - **Conditioned on covariates, each directed edge $(i,\, j)$ is generated by an independent Bernoulli trial with probability p_{ij}**
  - **Probability of connection is given by**

$$p_{ij} = \mathbb{E}\left(a_{ij}\right) = g\left(\sum_{k=1}^{p} \beta_k x_{ij}^{(k)}\right) = g\left(\beta_{k^*}\right)$$

Generalized Linear Models for Networks: Example

- **Note that for categorical covariates, estimates obtained in closed form as the log-odds**

$$\hat{\beta}_k = \log\left(\frac{\eta_k}{\zeta_k - \eta_k}\right)$$

where η_k is the number of observed edges in each category and ζ_k is the total number of possible edges in each category

- **Similarly, the estimate of the expected adjacency matrix can be expressed in a low-rank form**

$$\bar{\mathbf{A}} = \underbrace{\mathbf{C}}_{n\times c}\underbrace{\widehat{\mathbf{B}}}_{c\times c}\underbrace{\mathbf{C}^T}_{c\times n}$$

$$\mathbf{E} = \mathbf{A} - \bar{\mathbf{A}} = \mathbf{A} - \mathbf{C}\widehat{\mathbf{B}}\mathbf{C}^T$$

Exploitable Approximations

- **With the Chung–Lu model, the residuals matrix has a sparse-plus-rank-1 structure**
  - **This structure enables tractable computation of eigenvalues and eigenvectors**
- **In general, a GLM will not have such structure**
- **If probabilities are small, the logistic can be approximated as an exponential**
- **If the edge categories are coarse, this yields a low-rank structure for the probability matrix**
- **This allows the principal eigenspace to be computed for massive sizes**

$$p_{ij} = \frac{1}{1 + \exp(\underbrace{-x_{ij}^T \beta_{ij}}_{\text{edge category covariates}} \underbrace{- x_i^T \beta_i - x_j^T \beta_j}_{\text{vertex-specific covariates (source and destination)}})}$$

No obvious exploitable structure

$$p_{ij} \approx \frac{\exp(x_i^T \beta_i + x_j^T \beta_j)}{1 + \exp(-x_{ij}^T \beta_{ij})}$$

$$P \approx \underbrace{D(\exp(X^T \beta))}_{\text{diagonal}} \underbrace{\left\{ \frac{1}{1 + \exp(-x_{ij}^T \beta_{ij})} \right\}}_{\text{low rank}} \underbrace{D(\exp(X^T \beta))}_{\text{diagonal}}$$

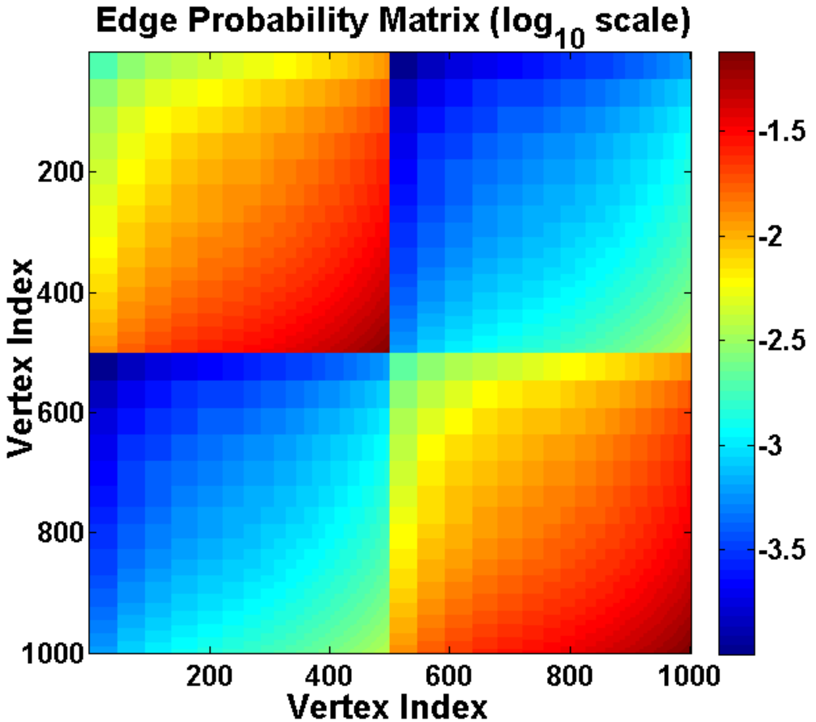
Approximation based on small probabilities enables analysis of principal eigenspace

Outline

- An anomaly detection framework for massive graphs
- Modeling attributed graphs using generalized linear models
- **Empirical results**

Empirical Example: Setup

- **10,000-trial Monte Carlo anomaly detection simulation**
- **For each trial, the observation is a 1,000-vertex graph**
- **Each graph is generated by a Chung–Lu/Stochastic Blockmodel hybrid**
  - Partitioned into two halves
  - Each half has higher probability of internal than external connectivity
  - Each vertex also has a "popularity" parameter
- **Two scenarios for embedded anomaly (8-vertex Erdős–Rényi graph)**
  - All 8 vertices on one side of the partition
  - 4 vertices on each side
- **Detection based on spectral norm of residuals matrix**



$$p_{ij} = \frac{1}{1+\exp(-\beta_{ij} - \beta_i - \beta_j)}$$

β_{ij}: dependent on whether *i* and *j* are both in the first half of the vertex set, both in the second half, or one in each

β_i, β_j: "popularity" parameter for individual vertices

Residuals Matrices

Given True Probabilities	Given Approximate Probabilities	Estimated Approximate Probabilities

$$p_{ij} = \frac{1}{1+\exp(-\beta_{ij}-\beta_i-\beta_j)}$$

- Use the matrix of Bernoulli parameters that generated the observed graph
- Demonstrates performance in an idealized situation

$$p_{ij} = \frac{\exp(\beta_i+\beta_j)}{1+\exp(-\beta_{ij})}$$

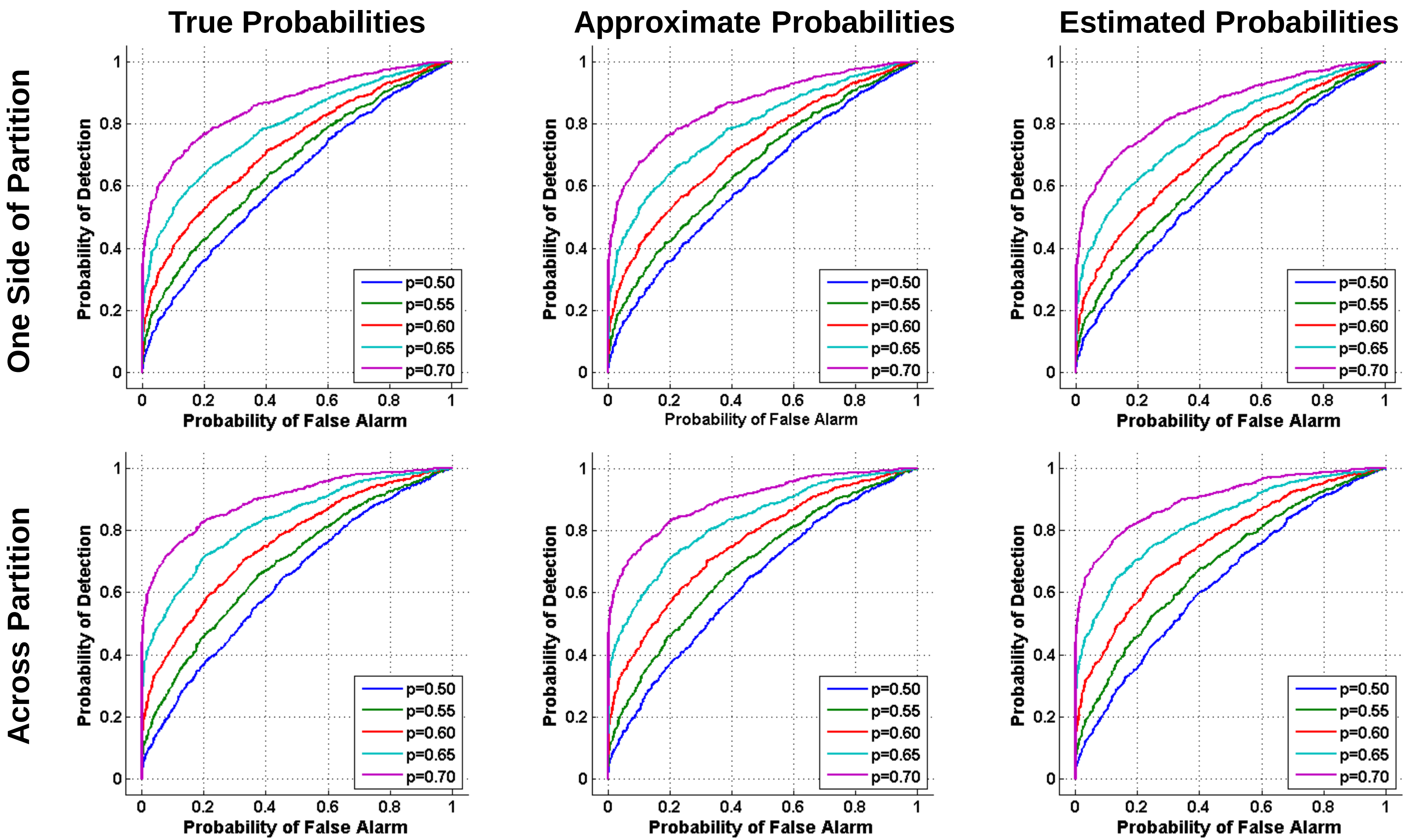
- Approximate probabilities are log-linear in popularity-based parameters
- Demonstrates the impact of using a computationally exploitable model

$$P = \begin{bmatrix} \hat{w}_1 & 0 \\ 0 & \hat{w}_2 \end{bmatrix} \begin{bmatrix} 1 & \hat{\alpha} \\ \hat{\alpha} & 1 \end{bmatrix} \begin{bmatrix} \hat{w}_1^T & 0 \\ 0 & \hat{w}_2^T \end{bmatrix}$$

- Probability matrix is estimated using a very simple estimator based on observed densities and degrees
- Demonstrates the loss in performance when not given model parameters

Use different residuals matrices to capture the effects of approximation and estimation

Detection Performance

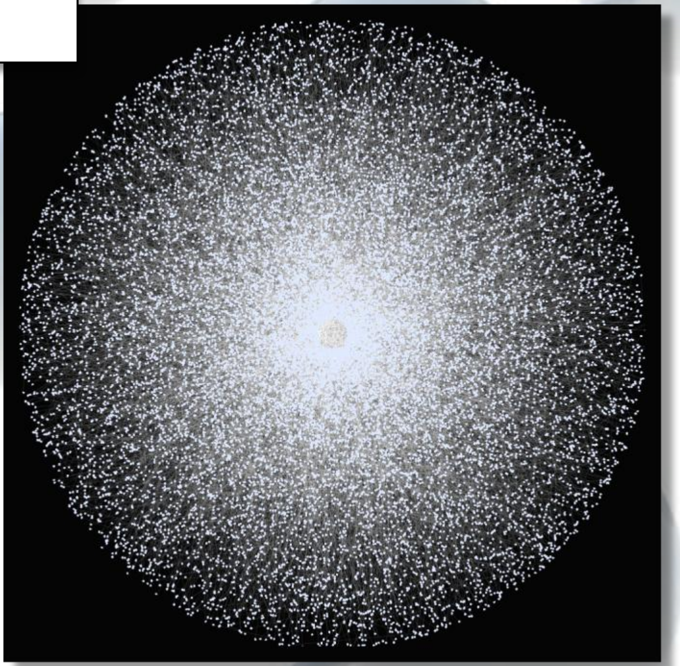


Computationally exploitable model yields nearly the same performance as true model

Thomson Reuters “Web of Science” Database

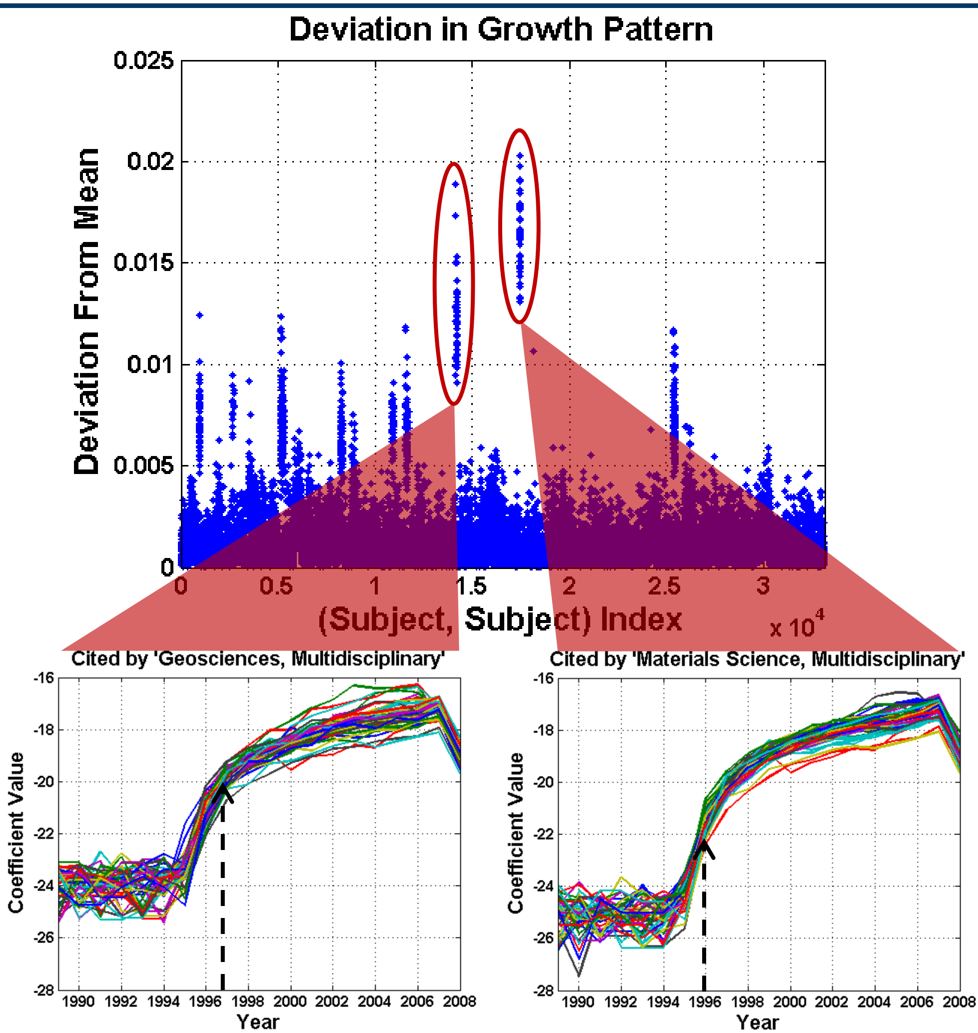
- **Citation database for papers in the sciences, social sciences, arts, and humanities**
  - **42 million records from 1900 to present**
  - **Articles from over 12,000 journals and 148,000 conference proceedings**

- **Records typically include**
  - **Author(s), title, publication date, type**
  - **Document IDs for works cited**
  - **May also include a number of other fields, e.g. subject area, institution, keywords, abstract**



Large Deviations in Subject Coefficients

- **Lower-dimensional residuals example: residuals of coefficients over time**
- **Consider growth patterns over time of coefficients for each subject–subject pair,**
- **Two blocks stand out significantly**
  - **One is citations by documents in the subject "Materials Science, Multidisciplinary"**
  - **The other: "Geosciences, Multidisciplinary"**
- **Both subjects were identified by about 100 more journals (including existing journals) in 1996 and 1997 than previous years**
- **For further analysis: Is this organic to the entities, or a collection artifact?**



Phase transition based in 1996/1997

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Summary

- **Anomaly detection framework for massive graphs**
  - **Residuals-based analysis for anomalous subgraph detection**
  - **Wish to incorporate side data (covariates) as attributed graphs**
  - **Need special structure to enable computational tractability**
- **Empirical results demonstrate use of GLMs and effectiveness of simplifying approximations in residuals analysis**
- **Future directions**
  - **Computationally tractable approaches to estimation and anomaly detection for more complex covariate structures**
  - **Effect of structural zeros and estimation of risk set**