Emergent Behavior Detection in Massive Graphs

Nadya T. Bliss and Benjamin A. Miller
MIT Lincoln Laboratory

15 February 2012

Minisymposium on Parallel Analysis of Massive Social Networks
SIAM Conference on Parallel Processing for Scientific Computing

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via 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.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA or the U.S. Government.
Large Scale Data & Information Processing

Novel Analytics:
Graph Analysis
Network Discovery

High Level Composable API:
D4M

Distributed Database:
Accumulo

High Performance Computing:
LLGrid

WEAK SIGNATURES,
NOISY DATA,
DYNAMICS

G = G_a + G_s
• background graph + foreground graph

High Performance Computing:
LLGrid

Interactive
super-computing

Distributed
database/
distributed file
system

Array algebra

Distributed Database:
Accumulo

High Level Composable API:
D4M

Novel Analytics:
Graph Analysis
Network Discovery
Graphs and Networks
From Data to Relationships

Graphs are a natural way to represent relationships between entities

**ISR**
- Graphs represent entities and relationships detected through multi-INT sources
- CONOPS: Identify anomalous patterns of activities

**Cyber**
- Graphs represent communication patterns of computers on a network
- CONOPS: Detect cyber attacks or malicious software

**Social**
- Graphs represent relationships between individuals or documents
- CONOPS: Identify hidden networks

Graph analytics enable detection of unobserved coordination between discrete elements or events in massive, noisy datasets
Efforts at MIT Lincoln Laboratory

Lincoln Laboratory is involved in several aspects of graph research; this presentation focuses on analytics.
Detecting Emergent Activity

Evolving planning of violent activity

Emerging coordinated activity (e.g., botnets)

Emerging technologies in scientific literature

Detection of emergent behavior is applicable to a variety of domains
Outline

• Introduction

• Detection Framework

• Example Dataset: Web of Science

• Ongoing Work: New Algorithms and Models

• Summary
Graph Based Residuals Analysis

Linear Regression

- Least-squares residuals from a best-fit line
- Analysis of variance (ANOVA) describes fit
- “Explained” vs “unexplained” variance → signal/noise discrimination

Graph “Regression”

- “Residuals” from a best-fit graph model
- Analysis of variance from expected topology
- Unexplained variance in graph residuals → subgraph detection
Subgraph Detection Algorithm Overview

Processing chain* for subgraph detection analogous to a traditional signal processing chain

**Input:**
- $A$, adjacency matrix representation of $G$
- No cue

**Output:**
- $v_s$, set of vertices identified as belonging to subgraph $G_f$

Residuals Construction

\[
B = A - \frac{KK^T}{2M}
\]

EXAMPLE:

GRAPH G

\[
\begin{pmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 \\
1 & * & * & * & * & * & * \\
2 & * & * & * & * & * & * \\
3 & * & * & * & * & * & * \\
4 & * & * & * & * & * & * \\
5 & * & * & * & * & * & * \\
6 & * & * & * & * & * & * \\
7 & * & * & * & * & * & * \\
\end{pmatrix}
\]

• The modularity matrix* is an example of a residuals matrix
  • Commonly used to evaluate quality of division of a graph into communities
  • Observed minus “expected” edges, yielding residuals

Eigen Decomposition

\[ B = UDU^T \]

Eigenvalues, sorted by magnitude

\[
\begin{bmatrix}
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_{N-1} \\
\lambda_N
\end{bmatrix}
\]

Corresponding eigenvectors

\[
\begin{bmatrix}
u_1 & u_2 & \ldots & u_{N-1} & u_N
\end{bmatrix}
\]

Principal components

Projection onto principal components of the modularity matrix yields good separation between background and a foreground with large residuals.

- Each point represents a vertex in \( G \)
- Vertices in \( G \): 1024
- Vertices in \( \hat{G} \): 12
- Uncued background/foreground separation

**12-vertex Subgraph in Powerlaw**

- background
- foreground

COLORING BASED ON KNOWN TRUTH

RESIDUALS CONSTRUCTION ➔ EIGEN DECOMPOSITION ➔ COMPONENT SELECTION ➔ DETECTION ➔ IDENTIFICATION
Toy Detection Example

Background (normal behavior), $G_b$ only

Background and foreground (includes anomalous behavior), $G_b + G_f$

$H_0$ and $H_1$ distributions are well separated

RESIDUALS CONSTRUCTION → EIGEN DECOMPOSITION → COMPONENT SELECTION → DETECTION → IDENTIFICATION

TEST STATISTIC: SYMMETRY OF THE PROJECTION ONTO SELECTED COMPONENTS

Powerlaw Background, 12-Vertex Dense Subgraph

Dynamic Graph Processing Chain

Input:
• $A(t)$, time-varying adjacency matrix representation of $G$
• No cue

Output:
• $v_s$, set of vertices identified as belonging to subgraph $G_f$

Processing chain for *dynamic* subgraph detection incorporates *temporal integration*
Temporal Integration

Temporal integration achieved by creating a linear combination of individual time-step graphs

\[
\tilde{B}(n) = \sum_{k=0}^{\ell-1} h(k) B(n-k)
\]

Weighted sum of residuals

Modularity (residuals) matrix at time \( n \)

Filter coefficients (scalars)
Outline

- Introduction
- Detection Framework
- Example Dataset: Web of Science
- Ongoing Work: New Algorithms and Models
- Summary
Thomson Reuters “Web of Science” Database

- Commercially available research database of papers in the sciences, social sciences, arts, and humanities
  - More than 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, as provided by the publication
Examples of Web of Science (WOS) Graphs

**CITATION GRAPH**
- nodes are papers
- (directed) edges are citations

**CO-AUTHOR GRAPH**
- nodes are authors
- (undirected) edges indicate co-authorship

**N-GRAM GRAPH**
- nodes are title subsequences
- (weighted) edges count co-occurrence

- Paper #2 cites paper #4
- A. PERSON and J. DOE have co-authored a paper together
- “STATISTICS OF” and “METHOD FOR” co-occur in 5,334 document titles
Eigenspace Analysis of Citation Graph

- 4,668,824 documents (documents in the database and those cited)
- Modularity matrices integrated over a 5-year window with a linear ramp filter
- Exceptionally large eigenvalues coincide with documents with thousands of citations (typically review articles)

Eigenvalues significantly larger than the general trend correspond to “clutter”-type behavior

Emerging Clusters: Citation Graph

- Consider vertices aligned with principal eigenvectors after manually pruning large citation lists
  - i.e., ignore eigenvectors highly concentrated on a single vertex
- Two subsets of vertices with significant internal connectivity
  - Most documents are in biochemistry and microbiology
  - Largely focus on metabolic properties of acids and proteins
- Temporal integration increases the strength of these subsets in the residuals space

Emerging interconnectedness in a major research area emphasized by linear ramp filter
Emerging Clusters: Coauthor Graph

• Consider vertices aligned with principal eigenvectors after manually pruning large author lists
  • Ignore eigenvectors aligned with sudden cliques
• Again, two tightly-connected subgraphs emerge
  • Similar periodicals in the (newly-founded) American J. Med comprise the other cluster
  • In both cases, teams of medical researchers form gradually over time
• Ramp filter emphasizes the densifying behavior

Emerging collaboration network stands out after temporal integration
Outline

• Introduction
• Detection Framework
• Example Dataset: Web of Science
• Ongoing Work: New Algorithms and Models
• Summary
Preferential Attachment with Memory

- Preferential attachment is a popular model for graph evolution
  - New nodes connect to existing ones with probability proportional to degree
  - Does not account for recency

- New model: predict new attachment rates by applying a finite impulse response filter to the sequence of recent connection counts

- Captures more variability in the WOS citation network than other models in the literature

New perspective on preferential attachment; integrating this technique into residuals analysis
Filtering Away Periodic Behavior

- In some applications, connections from autonomous nodes exhibit periodic behavior
  - Observed in web traffic data
- Model the expected graph as a linear combination of previous graphs to account for this effect
- Recent results indicate that this technique helps detect non-periodic coordinated activity

Moving average filter emphasizes aperiodic behavior and improves anomaly detection

\[
\min \left\| G(t) - \tilde{G}(t) \right\|_F
\]

\[
\tilde{G}(t) = \sum_{k=1}^{\ell} h_k G(t - k)
\]
Metadata-Based Modeling

- Topology-based modeling has significant practical limitations
  - Connection likelihood depends on more than degree

- Additional vertex or edge data can improve the expected value model
  - Used, e.g., in link prediction

- A recent statistical framework* models the probability of connection as a function of vertex and edge parameters:
  \[
  \text{logit} \left( P(v_i \rightarrow v_j) \right) = x_{ij}^T \beta
  \]

Summary

- Applying a dynamic subgraph detection framework to large graph data to pull emergent activity out of the background

- In the Web of Science document database, emergent clusters are found underneath some clutter activity in citation and coauthorship graphs using a simple expected value term
  - Integrating residuals over time emphasizes dynamic behavior (e.g., densification), enabling the detection of gradually-growing subnetworks

- New models for graph behavior incorporate dynamics of the topology and vertex and edge metadata into the graph’s expected value

- Ongoing work includes experimentation with the entire WOS corpus and integration of the new models into the analysis of residuals
Acknowledgements

• Nick Arcolano
• Michelle Beard
• Jeremy Kepner
• Matt Schmidt

• Bill Campbell
• Cliff Weinstein

• Patrick Wolfe