Anomaly, Event, and Fraud Detection in Large Network Datasets

Leman Akoglu
Stony Brook University
Dept. of Computer Science
leman@cs.stonybrook.edu

Christos Faloutsos
Carnegie Mellon University
School of Computer Science
christos@cs.cmu.edu

ABSTRACT

Detecting anomalies and events in data is a vital task, with numerous applications in security, finance, health care, law enforcement, and many others. While many techniques have been developed in past years for spotting outliers and anomalies in unstructured collections of multi-dimensional points, with graph data becoming ubiquitous, techniques for structured *graph* data have been of focus recently. As objects in graphs have long-range correlations, novel technology has been developed for abnormality detection in graph data.

The goal of this tutorial is to provide a general, comprehensive overview of the state-of-the-art methods for anomaly, event, and fraud detection in data represented as graphs. As a key contribution, we provide a thorough exploration of both data mining and machine learning algorithms for these detection tasks. We give a general framework for the algorithms, categorized under various settings: unsupervised vs. (semi-)supervised, for static vs. dynamic data. We focus on the scalability and effectiveness aspects of the methods, and highlight results on crucial real-world applications, including accounting fraud and opinion spam detection.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; E.1 [**Data Structures**]: Graphs and networks

Keywords

graph mining, anomaly detection, event detection, fraud

1. MOTIVATION AND OVERVIEW

When analyzing data, knowing what stands out in the data is often at least, or even more important and interesting than learning about its general structure. The branch of data mining concerned with discovering rare occurrences in datasets is called abnormality detection. This problem domain has numerous applications in security, finance, health care, law enforcement, and many others. In addition to revealing suspicious behavior, anomaly detection is vital for spotting rare events, such as rare diseases or side effects in medical domain.

To tackle the abnormality detection problem, many techniques have been developed in past years, especially for spotting outliers and anomalies in unstructured collections of multi-dimensional data points. On the other hand, graphs provide a powerful machinery for representing a wide range of data types in physical, biological, social, and information systems. As such, graph data (a.k.a. network, relational data) have become ubiquitous in the last decade. As a result, researchers have recently intensified their study of methods for anomaly detection in structured *graph* data.

Graph representation of datasets inherently impose long-range correlations among the data objects. For example in a reviewer-product review graph data, the extent a reviewer is fraudulent depends on what ratings s/he gave to which products, as well as how other reviewers rated the same products to an extent how trustful their ratings are, which in turn again depends on what other products they rated, and so on. As can be seen, due to this long-range correlations in graph datasets, detecting abnormalities in graph data is a significantly different task than that for points lying in a multi-dimensional feature space. In addition, the problem is challenging due to the large scale and dynamic nature of real-world graphs.

The main highlights of this tutorial are the following.

- We give a comprehensive overview of abnormality detection techniques for graph data for the *first* time
- We thoroughly explore techniques from both data mining (unsupervised, exploiting graph structure) and machine learning (semi-/supervised methods, employing relational learning).
- We put the abnormality (anomaly, event, fraud) detection methods under a unified lens, point out their connections and applications on diverse real-world tasks, e.g. accounting fraud, opinion spam, auction fraud.

2. TARGET AUDIENCE AND LEARNING OBJECTIVES

The target audience of this tutorial are researchers and practitioners who wish to know the most important techniques for outlier, anomaly, fraud, and event detection, with a focus on graph data. The tutorial would be of interest both to the data mining and machine learning community.

The audience is not expected to be familiar with the area, however the attendees should have basic knowledge on graph mining and machine learning. Through this tutorial, the participants will learn important techniques to attack the anomaly detection problem in data represented as graphs. The techniques are presented under changing settings; with or without ground truth data as well as for static or dynamically changing data.

WSDM'13, February 4–8, 2013, Rome, Italy.
ACM 978-1-4503-1869-3/13/02.

Part 0: **Introduction**

Part I: **Anomaly detection in static data**

(a) Overview: Outlier detection in clouds of data points

(b) Anomaly detection in *graph* data

   i. Anomalies in unlabeled graphs [3,13]

   ii. Anomalies in node-/edge-labeled graphs [4,8]

Part II: **Event detection in dynamic data**

(a) Overview: Event detection in time series of data

(b) Event detection in time series of *graph* data

   i. Graph-distance based methods [10]

   ii. Phase-transition based methods [2,12]

Part III: ***Graph*-based fraud detection**

(a) Theory and Algorithms: Introduction to relational learning [5,11,14],

(b) App.s: Relational learning for fraud detection

   i. Online auction fraud [9], opinion spam [1]

   ii. Accounting fraud risk [7], Web spam [6]

Table 1: Tutorial outline

3. TUTORIAL OUTLINE

The tutorial has 3 major parts, each of which is also organized into 2 main sections.

The first part focuses on anomaly detection methods for (a) collections of multi-dimensional data points, and (b) static graph data. The former is presented both for numerical and categorical data. Anomaly detection in graph data is covered for both unlabeled and labeled graphs, where each node and edge contains a label to identify its type.

The second part focuses on change detection approaches for (a) temporal data sequences, and (b) time-varying graph data. The former is categorized into instantaneous versus drifting change, after which specialized approaches for graphs based on edit distances and connectivity structure are presented.

The third part focuses on fraud detection using graph-based approaches. Firstly in (a), we give background on the theory and algorithms for relational learning. Later in (b) we demonstrate these algorithms in action, for diverse real-world fraud detection applications.

4. RELATION TO WSDM 2013 AREAS

Given the long list of applications and challenges they pose, abnormality detection is a very popular research topic. Not only the problems are mathematically interesting from a scientific point of view but, with their many applications in diverse fields, are also appealing from practitioners' point of view. As a result, the abnormality detection draws attention from many researchers in both industry and academia.

On the other hand, given the popularity of the topic, it is increasingly hard to keep up with the information overload arising from new approaches developed worldwide. To make the matter worse, the new contributions are dispersed into, although strongly related, fragments. That is, methods in this topic are grouped into outlier, anomaly, fraud, event, change, drift, fault detection separately. Our tutorial brings these concepts in one place and highlights their connections.

5. TUTORS' BIO AND EXPERTISE

LEMAN AKOGLU is an Assistant Professor at Stony Brook University, and received her Ph.D. from Carnegie Mellon University in 2012. She has won 2 "Best Paper" awards, published 15 refereed articles in major data mining venues, and is one of the inventors of 3 U.S. patents, filed by IBM T. J. Watson Research. Her research interests are in data mining, machine learning, and applied statistics with a focus on pattern mining, and anomaly and event detection in large dynamic data using graph mining and compression.

CHRISTOS FALOUTSOS is a Professor at Carnegie Mellon University. He has received the Research Contributions Award in ICDM 2006, the Innovations award in KDD 2010, 18 "best paper" awards, and several teaching awards. He has given over 30 tutorials and over 10 invited distinguished lectures. His research interests include data mining for graphs and streams, fractals, and database performance.

6. REFERENCES

[1] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in review networks. In *Technical Report CMU-CS-12-130*, 2012.

[2] L. Akoglu and C. Faloutsos. Event detection in time series of mobile communication graphs. In *Army Science Conference*, 2010.

[3] L. Akoglu, M. McGlohon, and C. Faloutsos. OddBall: Spotting anomalies in weighted graphs. In *PAKDD*, 2010.

[4] W. Eberle and L. B. Holder. Anomaly detection in data represented as graphs. *Intell. Data Anal.*, 11(6):663–689, 2007.

[5] L. Getoor, N. Friedman, D. Koller, A. Pfeffer, and B. Taskar. Probabilistic relational models. In *Intro. to Stat. Relational Learning.* MIT Press, 2007.

[6] Z. Gyogyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proc. VLDB*, 2004.

[7] M. McGlohon, S. Bay, M. G. Anderle, D. M. Steier, and C. Faloutsos. Snare: a link analytic system for graph labeling and risk detection. In *KDD*, pages 1265–1274, 2009.

[8] C. C. Noble and D. J. Cook. Graph-based anomaly detection. In *KDD*, pages 631–636, 2003.

[9] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos. Netprobe: a fast and scalable system for fraud detection in online auction networks. In *WWW*, 2007.

[10] B. Pincombe. Anomaly detection in time series of graphs using arma processes. *ASOR Bulletin.*, 24(4):2–10, 2005.

[11] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[12] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. Graphscope: parameter-free mining of large time-evolving graphs. In *KDD*, pages 687–696, 2007.

[13] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*, pages 418–425, 2005.

[14] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *UAI*, pages 485–492, 2002.