Abstract

How can we model users’ preferences? How do anomalies, fraud, and spam effect our models of normal users? How can we modify our models to catch fraudsters? In this tutorial we will answer these questions - connecting graph analysis tools for user behavior modeling to anomaly and fraud detection. In particular, we will focus on the application of subgraph analysis, label propagation, and latent factor models to static, evolving, and attributed graphs.

For each of these techniques we will give a brief explanation of the algorithms and the intuition behind them. We will then give examples of recent research using the techniques to model, understand and predict normal behavior. With this intuition for how these methods are applied to graphs and user behavior, we will focus on state-of-the-art research showing how the outcomes of these methods are effected by fraud, and how they have been used to catch fraudsters.

Perspective and Target Audience

Perspective: In this tutorial we focus on understanding anomaly and fraud detection through the lens of normal user behavior modeling. The data mining and machine learning communities have developed a plethora of models and methods for understanding user behavior. However, these methods generally assume that the behavior is that of real, honest people. On the other hand, fraud detection systems frequently use similar techniques as those used in modeling “normal” behavior, but are often framed as an independent problem. However, by focusing on the relations and intersections of the two perspectives we can gain a more complete understanding of the methods and hopefully inspire new research joining these two communities.

Target Audience: This tutorial is aimed at anyone interested in modeling and understanding user behavior, from data mining and machine learning researchers to practitioners from industry and government. For those new to the field, the tutorial will cover the necessary background material to understand these systems and will offer a concise, intuitive overview of the state-of-the-art. Additionally, the tutorial aims to offer a new perspective that will be valuable and interesting even for researchers with more experience in these domains. For those having worked in classic user behavior modeling, we will demonstrate how fraud can effect commonly-used models that expect normal behavior, with the hope that future models will directly account for fraud. For those having worked in fraud detection systems, we hope to inspire new research directions through connecting with recent developments in modeling “normal” behavior.
Outline

I. Introduction  (10 min)
   A. Graphs are a useful abstraction for a wide variety of domains: social networks, movie or product ratings and reviews, text in articles, medical diagnoses, financial transactions, etc.
   B. How can we make sense of the data? What does normal behavior look like? For example, we can predict ratings on Netflix or friends on Twitter and fill in missing information on Facebook.
   C. Fraud is rampant in nearly all of these applications - fake reviews on Yelp, purchased followers on Twitter, inflated trust on eBay, medical fraud, bank fraud. This activity deceives users and manipulates our prediction algorithms. Therefore it is important to understand how fraud influences our models and how we can isolate and catch anomalous behavior.

II. Subgraph Analysis and Patterns  (40 min)
   A. Background: Graph clustering and partitioning (5 min)
      i. Local search and graph queries [4, 19]
      ii. Co-clustering and cross associations [23, 13]
   B. Normal Behavior: Subgraph Patterns (10 min)
      i. Ego-nets: [29, 5]
      ii. Subgraph patterns in social networks: [40]
      iii. Influence of subgraphs on recommendation: [41]
   C. Abnormal Behavior: What are anomalous or fraudulent subgraphs? (25 min)
      i. Ego-net analysis: COI [14], OddBall [3]
      ii. Attributed subgraphs: FocusCO [35], SODA [20], CODA [16]
      iii. Temporal lockstep behavior: CopyCatch [8]; extended in [11]
      iv. Graph queries: “volcanoes” and “blackholes” on static graphs [30], attributed subgraph [46]
      v. Detecting fraud with co-clustering [34]
      vi. Graph cut for intrusion detection [15]

III. Label Propagation Methods  (1 hour)
   A. Random Walks and Eigenvectors (10 min)
      i. Background: Why do random walks “find” important parts of a graph?
   B. Belief and Label Propagation (10 min)
      i. Background: What is semi-supervised learning? What are belief and label propagation?
      ii. Normal Behavior: Predict attributes on nodes or why certain people are friends [12]
   C. Abnormal Behavior: How can random walks help us find fraud? (40 min)
      i. Surprising patterns in HITS: CatchSync [25]
      ii. Modifications of PageRank: TrustRank [22], SybilRank [10], CollusionRank [17], BadRank [43]
      iii. Use belief propagation for “guilt-by-association”:
         • Binary graphs: NetProbe [31]
         • Attributed graphs: FraudEagle [2], [26]
         • “Guilt-by-constellation” [42]
IV. Latent Factor Models (1 hour)

A. Background: What is the singular vector decomposition? (5 min)
   i. Generalization from Eigenvectors and HITS
   ii. Why do latent factor models, like SVD, work well for relational data?
   iii. What do the factors typically represent? Why? For example, the factorization of a user by movie ratings matrix gives genres; the decomposition of a word by document matrix gives topics.

B. Normal Behavior: Modifications for different settings: (15 min)
   i. Finding communities (binary matrices): MMSB [1], overlapping communities [14]
   ii. Missing data and prediction: SVD++ [28], BPMF [37], CoBaFi [7]
   iii. Multi-modal data: PARAFAC, Tensor factorization [33, 32]
   iv. Coupled factorization [39, 21]

C. Abnormal Behavior: What happens if there is fraud in the data? (40 min)
   i. Surprising patterns in latent factors (binary matrices): EigenSpokes [36], Get-the-scoop [24], fBox [38]
   ii. Surprising group patterns in ratings data: CoBaFi [7]
   iii. Surprising pattern in coupled factorization for heterogeneous graphs [18]
   iv. Group anomalies: GLAD [45]

V. Looking forward (10 min)

A. How can we handle multiple data sources and complex data?
B. With more complex data and methods, how can maintain the interpretability of discovered fraud?
C. Can we provide an adversarial analysis of our methods and offer provable guarantees?

Similar tutorials

- “What is Strange in Large Networks? Graph-based Irregularity and Fraud Detection” ([http://www3.cs.stonybrook.edu/~leman/icdm12/](http://www3.cs.stonybrook.edu/~leman/icdm12/)) ICDM 2012, Brussels, Belgium) - A subset of this tutorial focused on fraud detection systems, many of which are included in this tutorial. However, the tutorial was generally framed as finding “strange” structures and patterns in large graphs, some of which could be fraudulent patterns. This differs significantly from the perspective of this tutorial, which focuses on (1) the models and methods used in fraud detection systems and (2) their relation to modeling “normal” behavior.

Presenters

- Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation (1989), the Research Contributions Award in ICDM 2006, the Innovations award in KDD10, 20 “best paper” awards, and several teaching awards. He has served as a member of the executive committee of SIGKDD; he has published over 200 refereed articles, 11 book chapters and one monograph. He holds five patents and he has given over 30 tutorials and over 10 invited distinguished lectures. His research interests include data mining for graphs and streams, fractals, database performance, and indexing for multimedia and bio-informatics data. More details can be found at [http://www.cs.cmu.edu/~christos/](http://www.cs.cmu.edu/~christos/)
• **Leman Akoglu** is an Assistant Professor in the Department of Computer Science at Stony Brook University since August 2012. She received her Ph.D. from the Computer Science Department at Carnegie Mellon University. Her research interests span a wide range of data mining and machine learning topics with a focus on algorithmic problems arising in graph mining, pattern discovery, social and information networks, and especially anomaly, fraud, and event detection. Dr. Akoglu’s research has won 3 publication awards; one Best Knowledge Discovery Paper at ECML/PKDD 2009, a Best Paper at PAKDD 2010, and a Best Paper at ADC 2014. She also holds 3 U.S. patents filed by IBM T. J. Watson Research Labs. Dr. Akoglu is a recipient of the NSF CAREER award (2015) and an Army Research Office Young Investigator award (2013). Her research is supported by the National Science Foundation, the US Army Research Office, Office of Naval Research, and a gift from Northrop Grumman Aerospace Systems. More details can be found at [http://www.cs.stonybrook.edu/~leman](http://www.cs.stonybrook.edu/~leman).

• **Alex Beutel** is a fourth-year Ph.D. student at Carnegie Mellon University in the Computer Science Department. He previously received his B.S. from Duke University. His Ph.D. research focuses on large scale user behavior modeling, covering both recommendation systems and fraud detection systems. Alex’s research is supported by the National Science Foundation Graduate Research Fellowship Program and a Facebook Fellowship. More details can be found at [http://alexbeutel.com](http://alexbeutel.com).

References


[16] Jing Gao, Feng Liang, Wei Fan, Chi Wang, Yizhou Sun, and Jiawei Han. On community outliers and their efficient detection in information networks. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 813–822. ACM, 2010.


[46] Honglei Zhuang, Jing Zhang, George Brova, Jie Tang, Hasan Cam, Xifeng Yan, and Jiawei Han. Mining query-based subnetwork outliers in heterogeneous information networks.