Temporal Graph Mining for Fraud Detection

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

Pedro Fidalgo
Dept. of Information Science and Technology
University Institute of Lisbon
Lisbon, Portugal
pedro.fidalgo@mobileum.com

Mirela Cazzolato
Inst. of Mathematics and Computer Science
University of São Paulo (USP)
São Carlos, Brazil
mirelac@usp.br

***Abstract*— Given interacting entities (customers buying products; Twitter users re-tweeting posts; patients visiting doctors; machines sending packets to machines), how can we spot anomalies and possibly fraud? What should we do if we know that some entities are fraudulent?**

The problem has attracted huge interest; several suspicious patterns have been discovered, and several tools to spot such patterns have been developed. We survey the most successful, time-tested tools, starting with classic ones (matrix factorization, belief propagation, dense block detection) and continuing with newer, equally successful ones, and specifically multi-relational learning, and graph neural networks GNNs.

Focusing on practitioners, we also present some past success stories from diverse settings (phonecall networks, online retailers, social networks); we list some types of fraud and their tell-tale signs; and we also list the node-features that we found most useful (degree, core-number, etc), as well as some graph-visualization tools.

***Index Terms*—graph mining, fraud detection, visualization, time-evolving**

I. Rationale

Finding patterns and anomalies in graphs has numerous applications: in social networks, users may be re-tweeting posts after being paid to do so; in health-insurance fraud, patients collude with doctors to file fraudulent claims; in computer network security, machines sending packets to machines for a DDoS attack, and numerous more. The tutorial presents an organized list of time-tested tools, as well as fraud M.O.s (modus operandi), with emphasis on the intuition behind each tool and its applicability to real settings.

See section IV for relevance, and section VI for differences from earlier tutorials.

II. content details

1) **[5']** Introduction and Motivation. *[Fidalgo]*
2) **[10']** Types of fraud - 'behind enemy lines' *[Fidalgo]*
   a) in online retailers (Alibaba: brushing)
   b) in fake reviews (Flipkart)
   c) in Twitter (sock-puppets; reciprocity)
   d) in Facebook (synchronized behavior)
   e) in Weibo (sync, again)
   f) in phone networks (wan-giri, int-rev-share, int-bypass, camouflage; 'bridge' nodes from Pedro F.)
   g) in crypto (Srijan++)
3) **[25']** Static Graphs - unsupervised *[Faloutsos]*
   a) Patterns (degree distribution, conn. components, etc) Graph Mining book (Chakrabarti and Faloutsos) [1],
   b) Node Importance, Node Proximity, Link Prediction: SVD, PageRank [2], HITS [3], SALSA [4],
   c) Tools OddBall [5], CopyCatch [6], EigenSpokes [7], Fraudar [8]; Survey on anomaly detection [9]
   d) Communities/clusters METIS [10], Co-clustering [11], Cross-associations [12])
4) **[10']** Static Graphs - semi-supervised *[Faloutsos]*
   a) Featureless nodes: Belief propagation [13] and variants (FastBP [14] zooBP [15], [16])
   b) Applications/Success stories: NetProbe [17], Snare [18], Polonium [19].
5) **[10']** Time evolving graphs *[Faloutsos]*
   a) Patterns: spikeM [20], RSC [21], [22]
   b) Tools (tensors [23], timeCrunch [24]) AnomRank [25] SpotLight [26]
6) **[15']** General tools for outlier / micro-cluster detection *[Faloutsos]*
   a) Isolation Forest [27] and gen2Out [28]
   b) and many more: LOF [29]; LOCI [30]
   c) Even standard clustering methods: DBSCAN/OPTICS [31], [32]
7) **[5']** List of node-features (ever-growing) *[Cazzolato]*
   a) degree (in/out); weighted; unique-degree
   b) core#
   c) Inter-arrival time – IAT (median, mad, quantiles)
8) **[10']** Visualization tools *[Cazzolato]*
   a) Apring model [33]
   b) Adjacency matrix (after re-ordering) [34]
   c) 1-d distributions (pdf / ccdf / odds-ratio, in log-log)
   d) Scatter-plots / pair-plots (lookout) [35]
   e) Parallel axis [36]
   f) Demo of TgraphSpot from GitHub
9) **[5']** Conclusions *[Cazzolato]*

III. Target audience

Intended audience: Data scientists and practitioners, working on anomaly and fraud detection on large static and time-evolving graphs.

Thanks to the AIDA project (OCI-01-0247-FEDER-045907), co-funded by ERDF, COMPETE 2020 and FCT under CMU Portugal, and the São Paulo Research Foundation (FAPESP, grants No. 2016/17078-0, 2021/11403-5, 2020/11258-2).

Prerequisites: freshman matrix algebra (matrix multiplication, definition of eigenvalues), basic probability.

IV. Relevance and Rationale

See section I for the rationale.

V. List of forums

This is the first time we offer this tutorial. Some parts of it (static patterns, outline item 3, and semi-supervised methods, outline item 4) have been presented in graph mining tutorials (KDD17, WSDM13 - see below for the differences).

VI. List of past tutorials

- WSDM 2013 tutorial (*Anomaly, Event, and Fraud Detection in Large Graph Datasets*) proposal - and website with lecture slides. This is 10-year-ago tutorial - we have added recent material and recent papers.
- KDD 2017 tutorial (*Data Driven Approaches towards Malicious Behavior Modeling*) (slides). There was more emphasis on behavior modeling than graph mining and feature extraction.

VII. Tutors bios

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 SIGKDD Innovations Award (2010), the PAKDD Distinguished Contributions Award (2018), 31 "best paper" awards (including 8 "test of time" awards). He has given over 50 tutorials and over 25 invited distinguished lectures. His research interests include large-scale data mining with emphasis on graphs and time sequences; anomaly detection, tensors, and fractals.

E-mail: christos@cs.cmu.edu

Pedro Fidalgo: is the Engineering Director of Risk Management for R&D at Mobileum. He has a Msc in Computer Science from the University of Liverpool, UK and he is a PhD candidate in Complexity Sciences from (ISCTE-IUL), Portugal. His research interests include large-scale data mining with graphs, telecom fraud, anomaly detection and telecom signalling protocols.

E-mail: pedro.fidalgo@mobileum.com

Mirela Cazzolato: is a Postdoc Fellow in Computer Science at the Institute of Mathematics and Computer Science of the University of São Paulo (ICMC-USP) and the Heart Institute (InCor-USP), Brazil. She spent a year as a visiting researcher at Carnegie Mellon University (CMU), USA. She has a Ph.D. in Computer Science from ICMC-USP, with an internship period abroad at the Karlsruhe Institute of Technology (KIT), Germany. Her research interests include image analysis, graph mining, visualization, content-based retrieval, moving objects, and mHealth.

E-mail: mirelac@usp.br

VIII. Proposed length

2.5 hours.

IX. Slides/notes/videos

- Faloutsos: website with lecture slides; and keynote talk (York Univ., 2016)
- Cazzolato: TgraphSpot Presentation (BigData 2022); and TgrApp Demo Presentation (AAAI 2023);
- Fidalgo, Faloutsos, Cazzolato: AIDA Webinar.

References

[1] D. Chakrabarti and C. Faloutsos, *Graph Mining: Laws, Tools, and Case Studies*. Morgan Claypool, 2012.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford Digital Library Technologies Project, Tech. Rep., 1998, paper SIDL-WP-1999-0120 (version of 11/11/1999). [Online]. Available: http://dbpubs.stanford.edu/pub/1999-66

[3] J. Kleinberg, "Authoritative sources in a hyperlinked environment," in *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998, also appears as IBM Research Report RJ 10076, May 1997.

[4] R. Lempel and S. Moran, "SALSA: the stochastic approach for link-structure analysis," *ACM Trans. Inf. Syst.*, vol. 19, no. 2, pp. 131–160, 2001. [Online]. Available: http://doi.acm.org/10.1145/382979.383041

[5] L. Akoglu, M. McGlohon, and C. Faloutsos, "oddball: Spotting anomalies in weighted graphs," in *PAKDD (2)*, ser. Lecture Notes in Computer Science, vol. 6119. Springer, 2010, pp. 410–421.

[6] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, "Copycatch: stopping group attacks by spotting lockstep behavior in social networks," in *WWW*. International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 119–130.

[7] B. A. Prakash, M. Seshadri, A. Sridharan, S. Machiraju, and C. Faloutsos, "Eigenspokes: Surprising patterns and scalable community chipping in large graphs," in *ICDM Workshops*. IEEE Computer Society, 2009, pp. 290–295.

[8] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, "FRAUDAR: bounding graph fraud in the face of camouflage," in *KDD*. ACM, 2016, pp. 895–904.

[9] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data Min. Knowl. Discov.*, vol. 29, no. 3, pp. 626–688, 2015.

[10] G. Karypis and V. Kumar, "METIS: Unstructured graph partitioning and sparse matrix ordering system," *The University of Minnesota*, vol. 2, 1995.

[11] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Conference of the ACM Special Interest Group on Knowledge Discovery and Data Mining*. New York, NY: ACM Press, 2003.

[12] D. Chakrabarti, S. Papadimitriou, D. S. Modha, and C. Faloutsos, "Fully automatic Cross-associations," in *Conference of the ACM Special Interest Group on Knowledge Discovery and Data Mining*. New York, NY: ACM Press, 2004.

[13] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," in *NIPS*. MIT Press, 2000, pp. 689–695.

[14] D. Koutra, T. Ke, U. Kang, D. H. Chau, H. K. Pao, and C. Faloutsos, "Unifying guilt-by-association approaches: Theorems and fast algorithms," in *ECML/PKDD (2)*, ser. Lecture Notes in Computer Science, vol. 6912. Springer, 2011, pp. 245–260.

[15] D. Eswaran, S. Günnemann, and C. Faloutsos, "The power of certainty: A dirichlet-multinomial model for belief propagation," in *SDM*. SIAM, 2017, pp. 144–152.

[16] D. Eswaran, S. Günnemann, C. Faloutsos, D. Makhija, and M. Kumar, "Zoobp: Belief propagation for heterogeneous networks," *Proc. VLDB Endow.*, vol. 10, no. 5, pp. 625–636, 2017.

[17] 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*. New York, NY, USA: ACM, 2007, pp. 201–210.

[18] 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*. ACM, 2009, pp. 1265–1274.

[19] D. H. Chau, C. Nachenberg, J. Wilhelm, A. Wright, and C. Faloutsos, "Large scale graph mining and inference for malware detection," in *SDM*. SIAM / Omnipress, 2011, pp. 131–142.

[20] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa, “Fast mining and forecasting of complex time-stamped events,” in *KDD*. ACM, 2012, pp. 271–279.

[21] A. F. Costa, Y. Yamaguchi, A. J. M. Traina, C. T. Jr., and C. Faloutsos, “RSC: mining and modeling temporal activity in social media,” in *KDD*. ACM, 2015, pp. 269–278.

[22] C. Zang, P. Cui, and C. Faloutsos, “Beyond sigmoids: The nettide model for social network growth, and its applications,” in *KDD*. ACM, 2016, pp. 2015–2024.

[23] N. D. Sidiropoulos, L. D. Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, “Tensor decomposition for signal processing and machine learning,” *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, 2017.

[24] N. Shah, D. Koutra, T. Zou, B. Gallagher, and C. Faloutsos, “Timecrunch: Interpretable dynamic graph summarization,” in *KDD*. ACM, 2015, pp. 1055–1064.

[25] M. Yoon, B. Hooi, K. Shin, and C. Faloutsos, “Fast and accurate anomaly detection in dynamic graphs with a two-pronged approach,” in *KDD*. ACM, 2019, pp. 647–657.

[26] D. Eswaran, C. Faloutsos, S. Guha, and N. Mishra, “Spotlight: Detecting anomalies in streaming graphs,” in *KDD*. ACM, 2018, pp. 1378–1386.

[27] F. T. Liu, K. M. Ting, and Z. Zhou, “Isolation forest,” in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*. IEEE Computer Society, 2008, pp. 413–422. [Online]. Available: https://doi.org/10.1109/ICDM.2008.17

[28] M. Lee, S. Shekhar, C. Faloutsos, T. N. Hutson, and L. D. Iasemidis, “gen2out: Detecting and ranking generalized anomalies,” *CoRR*, vol. abs/2109.02704, 2021.

[29] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, “LOF: identifying density-based local outliers,” in *SIGMOD Conference*. ACM, 2000, pp. 93–104.

[30] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, “LOCI: fast outlier detection using the local correlation integral,” in *ICDE*. IEEE Computer Society, 2003, pp. 315–326.

[31] M. Ester, H. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, E. Simoudis, J. Han, and U. M. Fayyad, Eds. AAAI Press, 1996, pp. 226–231. [Online]. Available: http://www.aaai.org/Library/KDD/1996/kdd96-037.php

[32] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, “OPTICS: ordering points to identify the clustering structure,” in *SIGMOD Conference*. ACM Press, 1999, pp. 49–60.

[33] T. M. J. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” *Softw. Pract. Exp.*, vol. 21, no. 11, pp. 1129–1164, 1991.

[34] D. Chakrabarti, S. Papadimitriou, D. S. Modha, and C. Faloutsos, “Fully automatic cross-associations,” in *KDD*. ACM, 2004, pp. 79–88.

[35] N. Gupta, D. Eswaran, N. Shah, L. Akoglu, and C. Faloutsos, “Beyond outlier detection: Lookout for pictorial explanation,” in *ECML/PKDD (1)*, ser. Lecture Notes in Computer Science, vol. 11051. Springer, 2018, pp. 122–138.

[36] A. Inselberg, “The plane with parallel coordinates,” *Vis. Comput.*, vol. 1, no. 2, pp. 69–91, 1985.