Graph Mining and Multi-Relational Learning: Tools and Applications

Shobeir Fakhraei
Amazon

Christos Faloutsos
CMU and Amazon

November 28, 2020

Organizers

Shobeir Fakhraei is an Applied Scientist at CTPS Machine Learning Accelerator. Prior to Amazon, he has worked at various institutions including at University of Southern California, Microsoft Research, Yahoo! Labs, and University of California Santa Cruz, mainly researching and teaching applications of Machine Learning on Multi-Relational and Heterogeneous Graphs. He received his Ph.D. from the University of Maryland College Park on Statistical Relational Learning. He has published papers, been the program committee, and organized workshops at conferences such as KDD, ICML, NIPS, WWW, SDM, ICDM, and WSDM, including *KDD Mining and Learning with Graphs*[1] and *WSDM Heterogeneous Networks Analysis and Mining*[2] workshops.

E-mail: `shobeir@amazon.com`

Christos Faloutsos is a Professor at Carnegie Mellon University and an Amazon Scholar. 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), 28 "best paper" awards (including 7 "test of time" awards), he has given over 40 tutorials and over 20 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`

Abstract

Given a large graph, like who-buys-what, which is the most important node? How can we find communities? If the nodes have attributes (say, gender, or, eco-friendly, or fraudster), and we know the values of interest for a few nodes, how can we guess the attributes of the rest of the nodes?

[1] http://www.mlgworkshop.org/
[2] http://heteronam.org/

Graphs naturally represent a host of processes including interactions between people on social or communication networks, links between webpages on the World Wide Web, interactions between customers and products, relations between products, companies, and brands, relations between malicious accounts, and many others. In such scenarios, graphs that model real-world networks are typically heterogeneous, multi-modal, and multi-relational. With the availability of more varieties of interconnected structured and semi-structured data, the importance of leveraging the heterogeneous and multi-relational nature of networks in being able to effectively mine and learn this kind of data is becoming more evident.

In this proposal, we present time-tested graph mining algorithms (PageRank, HITS, Belief Propagation, METIS), as well as their connection to Multi-relational Learning methods. We cover both traditional, plain graphs, as well as heterogeneous, attributed graphs. Our emphasis is on the intuition behind these tools, with only pointers to the theorems behind them. The tutorial will includes many examples are from settings of direct interest to the Web Conference community (e.g., social networks, recommender systems, and knowledge graphs).

Topics

- Introduction and Motivation.
- Part 1: Plain Graphs - Traditional tools
  - 1.1: Node Importance, Node Proximity, Link Prediction: SVD, PageRank [1], HITS [2], SALSA [3],
  - 1.2 Community Detection METIS [4], Co-clustering [5], Cross-associations [6] 'No good cuts' [7])
  - 1.3: Fraud and Anomaly Detection OddBall [8], CopyCatch [9], EigenSpokes [10], Fraudar [11]; Survey on anomaly detection [12] Applications in Amazon: ClusterCatch
  - 1.4: Belief Propagation (Basic, FastBP, zooBP) [13]; FastBP [14] and extensions [15, 16]; Applications: NetProbe [17], Snare [18], Polonium [19]
- Part 2: Complex and Heterogeneous Graphs
  - 2.1: Factorization Methods: Factorization Machines [20, 21]; PARAFAC [22], Survey on tensors [23, 24], and applications [25, 26, 27]
  - 2.2: Heterogeneous Information Networks and Meta-path-based methods [28, 29]
  - 2.3: Multi-Relational and Statistical Relational Learning: Node Labeling [30], Link Prediction and Recommender Systems [31, 32, 33], Entity Resolution and Knowledge Graph Identification [34, 35]
- Conclusions

Relevance

Importance of leveraging the connectivity between objects of interest, and the heterogeneous and multi-relational nature of networks in being able to effectively mine and learn this kind of data is becoming more evident in many settings. This tutorial includes many examples from settings of direct interest to the Web Conference community (e.g., social networks, recommender systems, and knowledge graphs).

Duration

Three hours, evenly split among the two parts and the two presenters.

Interaction style

Lecture style.

Intended Audience and Level

Intended audience: Data scientists and practitioners, with interest on large graph, heterogeneous graphs, and tensor analysis.
Prerequisites: freshman matrix algebra (matrix multiplication, definition of eigenvalues), basic probability.
Learning Objectives: The participants will gain the intuition behind a set of classic and industry standard methods of graph mining, as well as 'recipes' on when to use them (and the rare cases on when not to). Moreover, they will obtain a quick, intuitive overview of multi-relational Learning methods, as well as applications of all these tools on real-world settings.

Previous Editions

The first part of the tutorial appeared in a KDD 2018 tutorial (https://www.cs.cmu.edu/~christos/TALKS/18-08-KDD-tut/)

Tutorial Materials

Attendees will be provided with the slides that contain working examples and many pointers to material for further learning about the presented topics. Presenters will provide necessary copyright permission to the organizers.

Online Format

The slides as well as the presenters' faces will be recorded during the tutorial presentation.

Video Snippet

- Video of distinguished lecture (Faloutsos, 2015, York University, Ontario Canada): https://www.youtube.com/watch?v=UyjhxEKjceA
- Video of research talk (Fakhraei, 2016, Allen Institute for AI): https://www.youtube.com/watch?v=izWoTtsMBIU

References

[1] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. Paper SIDL-WP-1999-0120 (version of 11/11/1999). 2

[2] Jon 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. 2

[3] Ronny Lempel and Shlomo Moran. SALSA: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, 2001. 2

[4] George Karypis and Vipin Kumar. Metis: unstructured graph partitioning and sparse matrix ordering system. Technical report, 1995. 2

[5] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic co-clustering. In *Conference of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, New York, NY, 2003. ACM Press. 2

[6] Deepayan Chakrabarti, Spiros Papadimitriou, Dharmendra S. Modha, and Christos Faloutsos. Fully automatic Cross-associations. In *Conference of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, New York, NY, 2004. ACM Press. 2

[7] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos. R-MAT: A recursive model for graph mining. *SIAM Int. Conf. on Data Mining*, April 2004. 2

[8] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. oddball: Spotting anomalies in weighted graphs. In *PAKDD (2)*, volume 6119 of *Lecture Notes in Computer Science*, pages 410–421. Springer, 2010. 2

[9] Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *WWW*, pages 119–130. International World Wide Web Conferences Steering Committee / ACM, 2013. 2

[10] B. Aditya Prakash, Mukund Seshadri, Ashwin Sridharan, Sridhar Machiraju, and Christos Faloutsos. Eigenspokes: Surprising patterns and scalable community chipping in large graphs. In *ICDM Workshops*, pages 290–295. IEEE Computer Society, 2009. 2

[11] Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. FRAUDAR: bounding graph fraud in the face of camouflage. In *KDD*, pages 895–904. ACM, 2016. 2

[12] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Min. Knowl. Discov.*, 29(3):626–688, 2015. 2

[13] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Generalized belief propagation. In *NIPS*, pages 689–695. MIT Press, 2000. 2

[14] Danai Koutra, Tai-You Ke, U. Kang, Duen Horng Chau, Hsing-Kuo Kenneth Pao, and Christos Faloutsos. Unifying guilt-by-association approaches: Theorems and fast algorithms.

In *ECML/PKDD (2)*, volume 6912 of *Lecture Notes in Computer Science*, pages 245–260. Springer, 2011. 2

[15] Dhivya Eswaran, Stephan Günnemann, and Christos Faloutsos. The power of certainty: A dirichlet-multinomial model for belief propagation. In *SDM*, pages 144–152. SIAM, 2017. 2

[16] Dhivya Eswaran, Stephan Günnemann, Christos Faloutsos, Disha Makhija, and Mohit Kumar. Zoobp: Belief propagation for heterogeneous networks. *PVLDB*, 10(5):625–636, 2017. 2

[17] Shashank Pandit, Duen Horng Chau, Samuel Wang, and Christos Faloutsos. Netprobe: a fast and scalable system for fraud detection in online auction networks. In *WWW*, pages 201–210, New York, NY, USA, 2007. ACM. 2

[18] Mary McGlohon, Stephen Bay, Markus G. Anderle, David M. Steier, and Christos Faloutsos. SNARE: a link analytic system for graph labeling and risk detection. In *KDD*, pages 1265–1274. ACM, 2009. 2

[19] Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, and Christos Faloutsos. Large scale graph mining and inference for malware detection. In *SDM*, pages 131–142. SIAM / Omnipress, 2011. 2

[20] Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000. IEEE, 2010. 2

[21] Steffen Rendle. Scaling factorization machines to relational data. In *Proceedings of the VLDB Endowment*, volume 6, pages 337–348. VLDB Endowment, 2013. 2

[22] Richard A Harshman and Margaret E Lundy. Parafac: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39–72, 1994. 2

[23] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. 2

[24] Nicholas D. Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Processing*, 65(13):3551–3582, 2017. 2

[25] Ching-Hao Mao, Chung-Jung Wu, Evangelos E. Papalexakis, Christos Faloutsos, Kuo-Chen Lee, and Tien-Cheu Kao. Malspot: Multi2 malicious network behavior patterns analysis. In *PAKDD (1)*, volume 8443 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2014. 2

[26] Ian N. Davidson, Sean Gilpin, Owen T. Carmichael, and Peter B. Walker. Network discovery via constrained tensor analysis of fmri data. In *KDD*, pages 194–202. ACM, 2013. 2

[27] Evrim Acar, Daniel M. Dunlavy, and Tamara G. Kolda. Link prediction on evolving data using matrix and tensor factorizations. In Yücel Saygin, Jeffrey Xu Yu, Hillol Kargupta, Wei Wang 0010, Sanjay Ranka, Philip S. Yu, and Xindong Wu, editors, *ICDM Workshops*, pages 262–269. IEEE Computer Society, 2009. 2

[28] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012. 2

[29] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*,

29(1):17–37, 2016. 2

[30] Shobeir Fakhraei, James Foulds, Madhusudana Shashanka, and Lise Getoor. Collective spammer detection in evolving multi-relational social networks. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 1769–1778, 2015. 2

[31] Pigi Kouki, Shobeir Fakhraei, James Foulds, Magdalini Eirinaki, and Lise Getoor. Hyper: A flexible and extensible probabilistic framework for hybrid recommender systems. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 99–106, 2015. 2

[32] Shobeir Fakhraei, Bert Huang, Louiqa Raschid, and Lise Getoor. Network-based drug-target interaction prediction with probabilistic soft logic. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(5):775–787, 2014. 2

[33] Shobeir Fakhraei, Dhanya Sridhar, Jay Pujara, and Lise Getoor. Adaptive neighborhood graph construction for inference in multi-relational networks. *arXiv preprint arXiv:1607.00474*, 2016. 2

[34] Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. Knowledge graph identification. In *International Semantic Web Conference*, pages 542–557. Springer, 2013. 2

[35] Pigi Kouki, Jay Pujara, Christopher Marcum, Laura Koehly, and Lise Getoor. Collective entity resolution in familial networks. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 227–236. IEEE, 2017. 2