Graph and Tensor Mining for Fun and Profit

Traditional-tutorial proposal for KDD'18

Xin Luna Dong
Amazon

Christos Faloutsos
SCS CMU

Andrey Kan
Amazon

Subhabrata Mukherjee
Amazon

Jun Ma
Amazon

Abstract

Given a large graph, which is the most important node? Can we plot and visualize the nodes in a low-dimensional space? Given a *heterogeneous* graph (where edges have attributes), like a knowledge graph, are there regularities? anomalies?

These questions and several related ones, have attracted huge interest, resulting in milestone algorithms like PageRank, HITS, recommendation systems, Belief Propagation, 'word2vec', and several more. This tutorial surveys all these algorithms, focusing on the *intuition* behind them (as opposed to the mathematical analysis); it highlights their strengths, similarities, and illustrates their applicability to real-world problems.

1. Title:

Graph and Tensor Mining for Fun and Profit

2. Abstract

Above

3. Target Audience and prerequisites

Data Scientists and practitioners, with interest on large graph analysis. Prerequisites: freshman matrix algebra (matrix multiplication, definition of eigenvalues).

4. Tutors

- Xin Luna Dong, Amazon, `lunadong@amazon.com`
- Christos Faloutsos, CMU and Amazon, +1-412-576.7932, `christos-sabbatical@cs.cmu.edu`
- Andrey Kan, Amazon, `avkan@amazon.com`

- Subhabrata Mukherjee, Amazon, subhomj@amazon.com
- Jun Ma, Amazon, junmaa@amazon.com

5. Tutors bio

The tutorial has contributions from all tutors, but it will be presented by Dr. Dong and Prof. Faloutsos.

Xin Luna Dong is a Principal Scientist at Amazon, leading the efforts of constructing Amazon Product Knowledge Graph. She was one of the major contributors to the Google Knowledge Vault project, and has led the Knowledge-based Trust project, which is called the "Google Truth Machine" by Washington Post. She has got the VLDB Early Career Research Contribution Award for "advancing the state of the art of knowledge fusion". She co-authored book "Big Data Integration", is the PC co-chair for Sigmod 2018 and WAIM 2015, and is serving in the VLDB advisory committee and the Board of Trustees of the VLDB Endowment. She has given several tutorials on data integration and knowledge management in top-tier conferences.

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), 25 "best paper" awards (including 5 "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.

Andrey Kan is an Applied Scientist at Amazon, working on estimating importance of nodes in Amazon Product Knowledge Graph. He has received his PhD from the University of Melbourne in 2013, and in the past three years, Andrey was an Instructor for Machine Learning subject in this university. Prior to joining Amazon, Andrey was developing mathematical models for biological applications. He was driving statistical analysis in projects published in Science, PNAS, and Nature Immunology, and his research was supported by multiple competitive research grants.

Subhabrata Mukherjee is a Machine Learning Scientist at Amazon building the Amazon Product Knowledge Graph. He is working on building large-scale machine learning models that extract knowledge from unstructured and semi-structured data. He graduated summa cum laude from Max Planck Institute for Informatics, Germany with a Ph.D. He has previously worked at IBM Research on domain adaptation of question-answering systems, and sentiment analysis. His research interests include probabilistic graphical models, information extraction, and recommender systems.

Jun Ma is an Applied Scientist at Amazon, working on knowledge graph embedding techniques for the Amazon Product Knowledge Graph. Before joining the team, he worked on fraud detection at Amazon using machine learning approaches. He has developed several innovative fraud detection methods such as latent factor models, large-scale graph algorithms, and deep sequence embedding. Jun received his Ph.D degree from Carnegie Mellon University in 2014, where he spent 6 years working on Monte Carlo methods in computational neuroscience.

6. Corresponding author with her/his email address

Christos Faloutsos: christos-sabbatical@cs.cmu.edu

7. Tutorial outline.

The intended duration is 3hours 30' total: 3h of presentation, and the rest, for questions.

- **[5']** Introduction - motivation.
- **[1h 25']** Part 1: Graphs
  - 1.1: properties of real graphs - power laws, shrinking diameters
  - 1.2: node importance (SVD, PageRank, HITS, SALSA, PLSA; link prediction / personalized PageRank)
  - 1.3 Community detection (METIS, co-clustering, 'no good cuts')
  - 1.4: Fraud/anomaly detection (oddBall, eigenspokes, copyCatch, Fraudar)
  - 1.5: Belief propagation (basic, FastBP, zooBP)
- **[1h 25']** Part 2: Tensors, KB (Knowledge Bases)
  - 2.1: basics: PARAFAC, Tucker, HAR
  - 2.2: embeddings (RESCAL, DistMult, ComplEx, TransE, NTN, R-GCN)
  - 2.3: inference (PRA, Fact Checking).
- **[5']** Conclusions

8. A list of earlier forums

The most related tutorial is 8 years ago:

- KDD 2009: *Large Graph-Mining: Power Tools and a Practitioner's Guide* (Faloutsos, Miller, Tsourakakis).
  https://www.cs.cmu.edu/~christos/TALKS/09-KDD-tutorial/
  In the current tutorial we emphasize recent developments; tools for anomaly detection; tensors and embeddings; and knowledge bases.

Other tutorials, with little overlap, are the following:

- Focusing on tensors exclusively.
  - SIGMOD 2007, *Mining Large Time-evolving Data Using Matrix and Tensor Tools*, Christos Faloutsos, Tamara G. Kolda, Jimeng Sun.
    https://www.cs.cmu.edu/~christos/TALKS/SIGMOD-07-tutorial/
- Focusing on node proximity and similarity.
  - ICDM 2014, *Node and graph similarity: theory and applications* ,
    Tina Eliassi-Rad, Danai Koutra, and Christos Faloutsos.
    https://web.eecs.umich.edu/~dkoutra/tut/icdm14.html
- Focusing on anomaly and fraud detection - no emphasis on tensors, knowledge bases, embeddings.
  - WSDM 2013, *Anomaly, Event, and Fraud Detection in Large Network Datasets*, Leman Akoglu and Christos Faloutsos
    https://www.andrew.cmu.edu/user/lakoglu/wsdm13/

- KDD 2015, *Graph-Based User Behavior Modeling: From Prediction to Fraud Detection*,
  Alex Beutel, Leman Akoglu and Christos Faloutsos.
  https://www.cs.cmu.edu/~abeutel/kdd2015_tutorial/
- KDD 2017, *Data-Driven Approaches towards Malicious Behavior Modeling*,
  Meng Jiang, Srijan Kumar, VS Subrahmanian, and Christos Faloutsos.
  http://www.meng-jiang.com/tutorial-kdd17.html

9. A list of the most important references that will be covered in the tutorial

- *Properties of graphs*: Graph Mining book (Chakrabarti and Faloutsos) [8], other patterns [2] [25].
- *Node importance and ranking*: PageRank [28], HITS [18], SALSA [23], PHITS [10], PLSA [16].
- *Belief Propagation* Yedidia et al [38]; FastBP [22] and extensions [14] [12] [13]; Applications: NetProbe [29], Snare [26], Polonium [9].
- *Anomaly/fraud detection*: OddBall [3], CopyCatch [6], EigenSpokes [32], Fraudar [17]; survey on anomaly detection [4].
- *Tensors* TopHITS [20] [21]; Comet communities [5]; Survey on tensors [19], [34]; Algorithms [39] [31] [30] and applications [24] [11] [1]
- *Embeddings*: bag of tricks [15], RESCAL [27], DistMult [37], ComplEx [36], TransE [7] NTN [35] R-GCN [33]

10. Equipment you will bring

Laptop; HDMI and VGA adaptors

11. Equipment you will need

- Projector, with HDMI or VGA input.
- Power sockets

12. Equipment attendees should bring

None

13-16. Hands-on-Tutorial

N/A

17. Slides

Slides will be available at `www.cs.cmu.edu` or `github`

18. Optional: Video snippet of you teaching

The tutorial (if accepted), will be delivered by Dr. Dong and Prof. Faloutsos.

- Xin Luna Dong, *Knowledge Vault and Knowledge-based Trust*, Stanford seminar series, 2015: *click here for video*.
- Christos Faloutsos, *Mining Large Graphs*, Distinguished Lecture Series, UIC, April 2015: *click here for the video* and *here for the foils*.

References

[1] E. Acar, D. M. Dunlavy, and T. G. Kolda. Link prediction on evolving data using matrix and tensor factorizations. In Y. Saygin, J. X. Yu, H. Kargupta, W. W. 0010, S. Ranka, P. S. Yu, and X. Wu, editors, *ICDM Workshops*, pages 262–269. IEEE Computer Society, 2009.

[2] L. Akoglu and C. Faloutsos. RTG: A recursive realistic graph generator using random typing. In *ECML/PKDD (1)*, volume 5781 of *Lecture Notes in Computer Science*, pages 13–28. Springer, 2009.

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

[4] L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Min. Knowl. Discov.*, 29(3):626–688, 2015.

[5] M. Araujo, S. Günnemann, S. Papadimitriou, C. Faloutsos, P. Basu, A. Swami, E. E. Papalexakis, and D. Koutra. Discovery of "comet" communities in temporal and labeled graphs com^2. *Knowl. Inf. Syst.*, 46(3):657–677, 2016.

[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*, pages 119–130. International World Wide Web Conferences Steering Committee / ACM, 2013.

[7] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.

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

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

[10] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *ICML*, pages 167–174. Morgan Kaufmann, 2000.

[11] I. N. Davidson, S. Gilpin, O. T. Carmichael, and P. B. Walker. Network discovery via constrained tensor analysis of fmri data. In *KDD*, pages 194–202. ACM, 2013.

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

[13] D. Eswaran, S. Günnemann, C. Faloutsos, D. Makhija, and M. Kumar. Zoobp: Belief propagation for heterogeneous networks. *PVLDB*, 10(5):625–636, 2017.

[14] W. Gatterbauer, S. Günnemann, D. Koutra, and C. Faloutsos. Linearized and single-pass belief propagation. *PVLDB*, 8(5):581–592, 2015.

[15] E. Grave, T. Mikolov, A. Joulin, and P. Bojanowski. Bag of tricks for efficient text classification. In *EACL (2)*, pages 427–431. Association for Computational Linguistics, 2017.

[16] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. 22nd SIGIR*, pages 50–57, 1999.

[17] 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*, pages 895–904. ACM, 2016.

[18] 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.

[19] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[20] T. G. Kolda, B. W. Bader, and J. P. Kenny. Higher-order web link analysis using multilinear algebra. In *ICDM*, pages 242–249, 2005.

[21] T. G. Kolda and J. Sun. Scalable tensor decompositions for multi-aspect data mining. In *ICDM*, pages 363–372. IEEE Computer Society, 2008.

[22] 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)*, volume 6912 of *Lecture Notes in Computer Science*, pages 245–260. Springer, 2011.

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

[24] C. Mao, C. Wu, E. E. Papalexakis, C. Faloutsos, K. Lee, and T. Kao. Malspot: Multi2 malicious network behavior patterns analysis. In *PAKDD (1)*, volume 8443 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2014.

[25] M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components: patterns and a generator. In *KDD*, pages 524–532. ACM, 2008.

[26] 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. ACM, 2009.

[27] M. Nickel, V. Tresp, and H. Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, pages 809–816. Omnipress, 2011.

[28] L. Page, S. Brin, R. Motwani, and T. 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).

[29] 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*, pages 201–210, New York, NY, USA, 2007. ACM.

[30] E. E. Papalexakis and C. Faloutsos. Fast efficient and scalable core consistency diagnostic for the parafac decomposition for big sparse tensors. In *ICASSP*, pages 5441–5445. IEEE, 2015.

[31] E. E. Papalexakis, C. Faloutsos, T. M. Mitchell, P. P. Talukdar, N. D. Sidiropoulos, and B. Murphy. Turbo-smt: Accelerating coupled sparse matrix-tensor factorizations by 200x. In *SDM*, pages 118–126. SIAM, 2014.

[32] 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*, pages 290–295. IEEE Computer Society, 2009.

[33] M. S. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. *CoRR*, abs/1703.06103, 2017.

[34] 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 Processing*, 65(13):3551–3582, 2017.

[35] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 926–934, 2013.

[36] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard. Complex embeddings for simple link prediction. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org, 2016.

[37] B. Yang, W. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. *CoRR*, abs/1412.6575, 2014.

[38] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *NIPS*, pages 689–695. MIT Press, 2000.

[39] S. Zhou, N. X. Vinh, J. Bailey, Y. Jia, and I. Davidson. Accelerating online CP decompositions for higher order tensors. In *KDD*, pages 1375–1384. ACM, 2016.