Detecting Communities and Anomalies in Large Real-world Graphs

Sungsu Lim
Dept. of Computer Science and Engineering
Chungnam National University
Daejeon, Korea
sungsu@cnu.ac.kr

Kijung Shin
Graduate School of AI & School of EE
KAIST
Daejeon, Korea
kijungs@kaist.ac.kr

***Abstract*—Graphs are a universal language for describing interactions in a variety of complex systems (e.g., the society, brains, and communication systems), and analyzing such graphs is crucial to understand and predict these complex systems. Communities (i.e., groups of densely-connected nodes) and anomalies, are key components of graphs, and identifying them in given graphs often leads to a much better understanding of the graphs.**

In this tutorial, we offer a comprehensive overview of the techniques for detecting communities and anomalies. Specifically, we first introduce basic and advanced techniques for detecting non-overlapping, overlapping, and multi-attribute communities, and then we introduce those for detecting anomalies based on features, spectral properties, and dense sub-structures. Additionally, we present a number of applications in a wide range of domains, including social media, e-commerce, and computer security.

I. Introduction and outline

Graphs are omnipresent, representing a wide range of data: social networks, web graphs, brain networks, to name a few. Thus, detecting underlying communities and anomalies in graphs is a vital task with numerous applications. For example, in brain networks, communities reveal functional organization of a brain, providing important intuition about its functions. In online social networks, subgraphs of unusual structure often indicate fraudulent behavior, such as fake followers on Twitter and fake 'Likes' on Facebook. Due to their importance, numerous algorithms have been developed for rapid and accurate identification of communities and/or anomalies. In this tutorial, we focus on providing a comprehensive overview of the state-of-the-art methods with their successful applications.

During the first half of this tutorial, we introduce some recent advances in detecting communities, i.e., groups of nodes that are densely connected internally and sparsely connected externally. We cover (1) basic clustering (or graph partitioning) [1]–[3], which divides a given graph into node-disjoint subgraphs, (2) overlapping clustering [4]–[7], which identifies cohesive subgraphs allowing mixed memberships for nodes, and (3) multi-attribute clustering [8]–[10], which finds meaningful communities of graphs with multiple attributes. Overall, we introduce various clustering methods with applications.

During the second half, we present the recent techniques for detecting anomalies, specifically unusual nodes, subgraphs, and changes in graphs. We cover (1) features-based methods [11]–[13], which exploit graph-centric features (e.g., degree, coreness, and triangle counts), (2) spectral-based methods [14]–[16], which detect anomalies after projecting nodes into spectral subspaces, and (3) density-based methods [17]–[21], which identify unusually dense sub-structures often resulted from synchronized behavior of anomalous nodes. In addition to their technical details, we present interesting anomalies identified by the methods from various real-world graphs.

A brief outline of the tutorial is as follows:

- Part I: Introduction
- Part II: Community Detection in Graphs
  1) Non-overlapping clustering [1]–[3]
  2) Overlapping clustering [4]–[7]
  3) Multi-attribute clustering [8]–[10]
- Part III: Anomaly detection in graphs
  1) Feature-based approaches [11]–[13]
  2) Spectral-based approaches [14]–[16]
  3) Density-based approaches [17]–[21]
- Part IV: Conclusions

II. Target Audience

This tutorial is targeted at anyone interested in data mining for graphs, from researchers to practitioners from industry. For the audience new to this field, the tutorial will cover necessary preliminaries and provide an intuitive overview of recent studies on community detection and anomaly detection on graphs. The tutorial will also offer in-depth descriptions of state-of-the-art techniques for the audience with more experience in this field.

III. Tutors' Bio and Experience

Sungsu Lim is an Assistant Professor in the Department of Computer Science and Engineering at Chungnam National University. He is also an advisor of NOTA Incorporated. He received his Ph.D. from the Graduate School of Knowledge Service Engineering at KAIST in 2016. He won the Qualcomm Innovation Award 2016. His research interests include mining and modeling large-scale social and information networks. More details can be found at http://sungsulim.com

Kijung Shin is an Assistant Professor in the Graduate School of AI and the School of EE at KAIST. He received his Ph.D. from the Computer Science Department at Carnegie

Mellon University in 2019. He has published 25 referred articles in major data mining venues, and he won the best research paper award at KDD 2016. His research interests span a wide range of graph mining topics, with a focus on designing scalable algorithms for large-scale graphs. More details can be found at http://kijungshin.com

References

[1] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, vol. 42, no. 1, pp. 181–213, 2015.

[2] S. Lim, J. Kim, and J.-G. Lee, "Blackhole: Robust community detection inspired by graph drawing," in *ICDE*, 2016.

[3] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2015.

[4] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in *WSDM*, 2013.

[5] J. J. Whang, D. F. Gleich, and I. S. Dhillon, "Overlapping community detection using neighborhood-inflated seed expansion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 5, pp. 1272–1284, 2016.

[6] A. Epasto, S. Lattanzi, and R. P. Leme, "Ego-splitting framework: from non-overlapping to overlapping clusters," in *KDD*, 2017.

[7] J. Kim, S. Lim, J.-G. Lee, and B. S. Lee, "Linkblackhole*: Robust overlapping community detection using link embedding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 11, pp. 2138–2150, 2019.

[8] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *ICDM*, 2013.

[9] J. Kim, J.-G. Lee, and S. Lim, "Differential flattening: A novel framework for community detection in multi-layer graphs," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 2, pp. 27:1–27:23, 2017.

[10] C. Zhe, A. Sun, and X. Xiao, "Community detection on large complex attribute network," in *KDD*, 2019.

[11] L. Akoglu, M. McGlohon, and C. Faloutsos, "Oddball: spotting anomalies in weighted graphs," in *PAKDD*, 2010.

[12] K. Henderson, B. Gallagher, L. Li, L. Akoglu, T. Eliassi-Rad, H. Tong, and C. Faloutsos, "It's who you know: graph mining using recursive structural features," in *KDD*, 2011.

[13] K. Shin, T. Eliassi-Rad, and C. Faloutsos, "Patterns and anomalies in k-cores of real-world graphs with applications," *Knowledge and Information Systems*, vol. 54, no. 3, pp. 677–710, 2018.

[14] B. A. Prakash, A. Sridharan, M. Seshadri, S. Machiraju, and C. Faloutsos, "Eigenspokes: Surprising patterns and scalable community chipping in large graphs," in *PAKDD*, 2010.

[15] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang, "Inferring strange behavior from connectivity pattern in social networks," in *PAKDD*, 2014.

[16] N. Shah, A. Beutel, B. Gallagher, and C. Faloutsos, "Spotting suspicious link behavior with fbox: An adversarial perspective," in *ICDM*, 2014.

[17] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, "Copycatch: stopping group attacks by spotting lockstep behavior in social networks," in *WWW*, 2013.

[18] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos, "Spotting suspicious behaviors in multimodal data: A general metric and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2187–2200, 2016.

[19] B. Hooi, K. Shin, H. A. Song, A. Beutel, N. Shah, and C. Faloutsos, "Graph-based fraud detection in the face of camouflage," *ACM Transactions on Knowledge Discovery from Data*, vol. 11, no. 4, p. 44, 2017.

[20] K. Shin, B. Hooi, and C. Faloutsos, "Fast, accurate and flexible algorithms for dense subtensor mining," *ACM Transactions on Knowledge Discovery from Data*, vol. 12, no. 3, pp. 28:1–28:30, 2018.

[21] K. Shin, B. Hooi, J. Kim, and C. Faloutsos, "Densealert: Incremental dense-subtensor detection in tensor streams," in *KDD*, 2017.