Leveraging Propagation for Data Mining: Models, Algorithms and Applications

B. Aditya Prakash
Department of Computer Science,
Virginia Tech,
Blacksburg, VA, USA
badityap@cs.vt.edu

Naren Ramakrishnan
Department of Computer Science,
Virginia Tech,
Arlington, VA, USA
naren@cs.vt.edu

ABSTRACT

Can we infer if a user is sick from her tweet? How do opinions get formed in online forums? Which people should we immunize to prevent an epidemic as fast as possible? How do we quickly zoom out of a graph? Graphs—also known as networks—are powerful tools for modeling processes and situations of interest in real life domains of social systems, cyber-security, epidemiology, and biology. They are ubiquitous, from online social networks, gene-regulatory networks, to router graphs.

This tutorial will cover recent and state-of-the-art research on how propagation-like processes can help big-data mining specifically involving large networks and time-series, algorithms behind network problems, and their practical applications in various diverse settings. Topics include diffusion and virus propagation in networks, anomaly and outbreak detection, event prediction and connections with work in public health, the web and online media, social sciences, humanities, and cyber-security.

Keywords

Graph Mining; Propagation; Dynamical Processes; Diffusion; Cyber Security; Public Health; Social Media

1. TARGET AUDIENCE, RELEVANCE, AND PREREQUISITES

Target Audience: The target audience is data mining, data management, data science researchers and practitioners in both academia and industry, who desire to learn more about models and tools based on propagation-like processes involving large datasets. There will be special emphasis on the cross-disciplinary aspect of the concepts and tools involved.

Relevance: The topic is very interdisciplinary, with connections to many high-impact areas—public health and epidemiology, systems biology, social sciences and humanities, cyber security, viral marketing and social media—hence this tutorial would be of interest to a broad cross-section of the KDD audience. Some of the numerous applications include disease modeling, trend forecasting, information cascade prediction, protest forecasting, data visualization, recommendation systems, social networks, and data compression. Due to the explosion in the availablity of large datasets, the emphasis of this tutorial will be on recent progress spanning novel algorithms, techniques, and new applications.

Prerequisites: For maximum benefit, the expected prerequisite is an undergraduate degree in computer science or a related field. However, the tutorial's emphasis is on the intuition behind the results and tools so that data science researchers and practitioners can broadly digest the concepts.

2. OUTLINE OF TUTORIAL

The tutorial consists of three parts overall (the intended total duration is 3hrs, with a 0.5hr break). The tutorial webpage is at: http://www.cs.vt.edu/~badityap/TALKS/16-kdd-tutorial/.

2.1 Fundamental Models [45min]

In this part, the goal is to understand various fundamental propagation models and learn about properties of such processes (like epidemic thresholds), in a variety of settings, e.g., static networks, single viruses, dynamic networks, and multiple competing contagions. We will also demonstrate how many such processes share underlying similarities core problems. We will particularly review recent analytical results based on these models and their practical implications.

The flow will be:

- Fundamental propagation models and relations to each other (homogenous or random graphs)
- Extension to arbitrary networks
- Tipping points and phase transitions in models
- Extensions to dynamic networks, competing viruses, and multi-profile networks

2.2 Algorithms [1h]

In this part of the tutorial we will focus on state-of-the-art inference and optimization algorithms for multiple settings introduced in Part 1 such as: immunization, influence estimation, cascade network inference, influence maximization, and reverse engineering epidemics. The focus will be on how to design algorithms on general arbitrary networks, with effective guarantees and which scale up to networks of

KDD '16 August 13-17, 2016, San Francisco, CA, USA

ACM ISBN 978-1-4503-4232-2/16/08.
DOI: http://dx.doi.org/10.1145/2939672.2945390

millions and billions of nodes. The algorithms use a variety of techniques including stochastic optimization, submodular optimization, survival theory, and the minimum description length principle.

The flow will be:

- Estimating and summarizing influence
- Immunization algorithms
- Influence maximization algorithms
- Diffusion network inference algorithms
- Finding culprits

2.3 Applications [1h 15mins]

In this part, we will focus on numerous applications, involving the use of propagation to solve different problems in various domains from graph mining, meme-tracking, tweets trends, predicting malware attacks (cyber-security), civil unrest forecasting, disease forecasting, spatio-temporal forecasting, and filling-in missing data. Here we also intend to show demos of some of our tools such as graph summarization and flu-trend prediction.

The flow of this part will be:

- General Graph mining
- Memes, Tweets, and Blog cascades
- Malware attacks
- Protest forecasting
- Disease trends
- Short demos (time permitting)

3. INSTRUCTORS

B. Aditya Prakash is an Assistant Professor in the Computer Science Department at Virginia Tech. He graduated with a Ph.D. from the Computer Science Department at Carnegie Mellon University in 2012, and received his B.Tech (in CS) from the Indian Institute of Technology (IIT) – Bombay in 2007. He has published more than 40 refereed papers in major venues, holds two U.S. patents and has given two tutorials (at VLDB 2012 and ECML/PKDD 2012). His work has also received a best paper award and two best-of-conference selections (CIKM 2012, ICDM 2012, ICDM 2011) and multiple travel awards. His research interests include Data Mining, Applied Machine Learning and Databases, with emphasis on big-data problems in large real-world networks and time-series. His work has been funded through grants/gifts from the National Science Foundation (NSF), the Department of Energy (DoE), the National Security Agency (NSA), the National Endowment for Humanities (NEH) and from companies like Symantec. He received a Facebook Faculty Gift Award in 2015. He is also a core faculty member at the Discovery Analytics Center at Virginia Tech. His homepage is at: http://www.cs.vt.edu/˜badityap.

Naren Ramakrishnan is the Thomas L. Phillips Professor of Engineering and Director of the Discovery Analytics Center at Virginia Tech. His research interests focus on data mining for intelligence analysis, forecasting, sustainability, and health informatics. He currently leads the IARPA OSI EMBERS project on forecasting critical societal events (disease outbreaks, civil unrest, and elections) using open source indicators. His research has been supported by NSF, DHS, NIH, NEH, IARPA, DARPA, DTRA, ONR, General Motors, HP Labs, and NEC Labs. Ramakrishnan serves on the editorial boards of ACM Transactions on Knowledge Discovery from Data, IEEE Computer, Data Mining and Knowledge Discovery, IEEE Transactions on Knowledge and Data Engineering, and other journals. He was an invited co-organizer of the National Academy of Engineering Frontiers of Engineering symposium in 2009. Ramakrishnan is an ACM Distinguished Scientist. His homepage is at: http://www.cs.vt.edu/˜naren.

Acknowledgements

This material is based on work partially supported by the National Science Foundation under grant no. IIS-1353346, the Maryland Procurement Office under contract H98230-14-C-0127, the Intelligence Advanced Research Projects Activity via DoI/NBC contract number D12PC000337, the National Endowment for Humanities under grant no. HG-229283-15, the Oak Ridge National Laboratory under Task Order 4000143330, the VT College of Engineering and a Facebook Faculty Gift. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the respective funding parties.