Learn how to see. Realize that everything connects to everything else.

–Leonardo da Vinci

MINING CONNECTIONS

Connections are ubiquitous in different domains; from web pages tangled with hyperlinks, and employees communicating by emails, to proteins and genes interacting to carry out a biological function, and neurons passing signals through synapses. My general research interest is studying ***interconnected data*** to draw insights from the structure and dynamics of complex systems. My short-term goal is to design models and algorithms that integrate ***three principal elements*** in formulating and solving real-world problems: ***connection***, ***content***, and ***time***.

Content and ***connection*** are complementary information sources. For example, when investigating human trafficking activity in online escort advertisement, we want to consider indicators extracted from the content of each ad, e.g., references to an underage victim; as well as how different ads are connected to each other, e.g., using a shared phone number, which may reveal an organized activity.

Time is orthogonal to content and connection, as they both change over time. Moreover, connections act as the infrastructure for diffusion processes. For example, consider modeling spread of hospital-acquired infections. Given the patient records, which includes their room assignments, the procedures they underwent, and their culture results, we can infer the potential infection propagation routes, e.g., being in the same room at the same time. These routes form a *dynamic structure*, over which we want to further model the *dynamic process* of spread of a disease.

APPROACH

My research methodology draws on ideas from network science, machine learning, and data mining to aggregate connection, content, and time. Machine learning techniques mainly focus on ***features of individual datapoints***, whereas network science methods primarily study the ***connections between them***. There exist multiple lines of research in between, notably *statistical relational learning* subsumes connections as relational rules, and *heterogeneous graph mining* augments the graph topology with meta-data for nodes and edges. However, substantial effort is still needed to enable expressive, general, scalable, and time-dependent modeling for the complex data we encounter in real-world settings.

As a well-studied example, consider clustering the webpages in the world-wide-web. We can use a topic modeling algorithm to group the webpages based on the similarity of the terms they use (*content*); or utilize a random walk on their hyperlinks (*connection*) to find closely related webpages, within which the random walk is more likely to be trapped. To further account for how the web evolves (*time*), we can devise streaming and incremental [23] extensions which examine how these groups merge and split over time [24], and how different webpages move between the groups [1]. Drawing together these approaches is not straightforward [18, 22]. In the past, I have extensively worked on transferring the well-studied clustering evaluation practices to network science. In an earlier work, we reformulated the commonly-used clustering criteria in terms of the connections [12, 13]; which became an encyclopedia article [14]. In a subsequent effort, we introduced an agreement measure for a pair of clustering which can incorporate the structure [19]; see Figure 1. Our

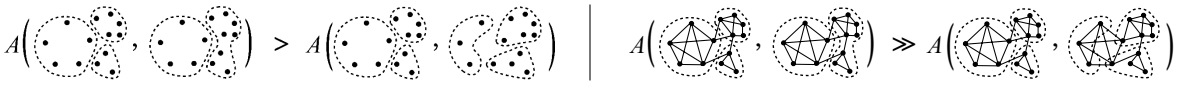


Figure 1: Left: A clustering agreement measure compares the pair-wise similarity of two clusterings of a same dataset, often with the ground-truth clustering. Right: Including the connections between the data-points makes the agreement of the first pair of clusterings stronger, since it is better aligned with the structure of the data.

formulation further generalizes the current widely-used measures and provides elegant extensions for when groups could have overlaps [20].

My ongoing effort towards richer models for interconnected data is threefold: exploration, unification, and expansion. ***Exploration*** studies real-world data to recognize the inherent patterns, in particular by investigating the interplay between the connections and content. ***Unification*** redefines different classical problems for interconnected data. For example, we are formulating structural-dependent component analysis, and joint representation learning for attributed graphs, which tie to multiple network science tasks: including link and attribute prediction, graph summarization, node classification, and anomaly detection. ***Expansion*** devises new tasks which arise when solving real-world problems, such as active link inference discussed later. In the following, I elaborate on these three directions within the context of their broader impacts.

BROADER IMPACTS

COMPUTATIONAL SOCIAL SCIENCE: Social networking sites are changing societies in ways we do not have the tools to comprehend yet. Models which encompass content, connection, and time enable us to gain insights into the complex dynamics of social networks – by bridging between psychology, i.e., the study of individuals, and sociology, i.e., the study of relationships.

To this end, we recently looked at how interests of users (content) interrelates with their friendships (connections) across multiple social networking sites. We developed a novel index to measure the *structural correlation* between their interests [10, 11], and then studied ***universal patterns*** which go beyond correlation [2]; see Figure 2. One of my short-term research goals is to extend the coverage of this work, explore other universal patterns, and apply the findings to develop new techniques for characterization, modelling, and anomaly detection in these networks. In the long term, I am interested in developing computational techniques for understanding the structure and dynamics of online societies.

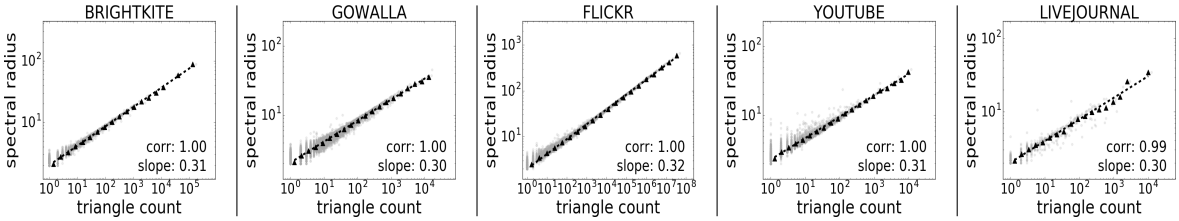
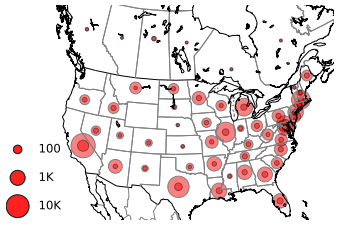


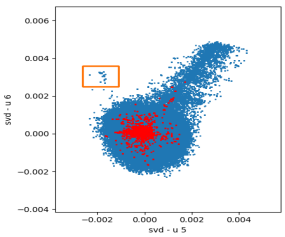
Figure 2: An example pattern, power-law relationship between the spectral radius (highest singular value of the adjacency matrix) and triangle count (number of closed triads), observed across five social networking sites. Each point corresponds to the connections between the users with a particular interest, e.g., Starbucks goers.

COMBATING ONLINE HUMAN TRAFFICKING: Online advertising market provides an easy to use and low-risk platform for human traffickers, allowing them to remain anonymous and have broader geographic operations. Escort advertisements are growing at unprecedented scale, already creating hundreds of millions in revenue, while a majority of sex-trafficking victims report of being advertised online. Part of my research program is focused on developing data-driven techniques to provide actionable intelligence to counter human trafficking.

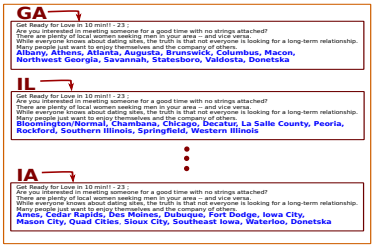
We recently developed an ***active link inference*** method which helps an investigator to "connect the dots" and spot an ***organized activity*** within millions of online escort advertisements [6, 7]. Following this, we are currently investigating different ***anomaly detection*** methods, which take a bird's eye view of this intertwined data to recognize different posting patterns, and spot anomalous and suspicious behavior. We are in particular exploring recent techniques in learning low-dimensional embeddings for nodes in a graph using deep neural networks and comparing them with the classic spectral subspace embeddings.



(a) **A wide-spread issue**: reported human trafficking cases.



(b) **A spectral subspace**: isolates a cluster of related ads.



(c) **Discovered nationwide activity**: similar ads in cities of different states.

Figure 3: Figure 3a shows the spatial distribution of online escort advertisements using a phone number associated to a human trafficker, posted only in a three months period on one of the primary hubs for escort ads. Figure 3b presents a spectral embedding of about 4 million advertisements from the same website. Figure 3c shows the three advertisements from the spotted microcluster of Figure 3b which exhibit suspicious behavior.

EDUCATIONAL DATA MINING: There is a growing number of courses delivered using e-learning environments. Discussion forums within these environments provide the basis for collaborative learning. Manually evaluating the participation of students in these forums is a challenge, considering that current systems do not provide the necessary tools for such analysis.

In a previous work, we demonstrated the usefulness of network science techniques in monitoring the ***engagement level of students*** in online courses [15, 9]. We modeled the interactions between the students and the terms used in their communications as two coupled networks which are both automatically extracted from the discussion forums. We then analyzed these networks to rank the participation of students, group the collaborating students, and organize the terms used in their discussion topics. Our analysis enables the instructor to spot changes and make proper recommendation to the students, throughout the course. This work was featured in the SIGKDD Explorations newsletter [16]. Proper analysis of educational data could have a substantial effect on the effective delivery of online courses and shape the future of education. Motivated by this impact, a part of my future research plan is to contribute to educational data analysis.

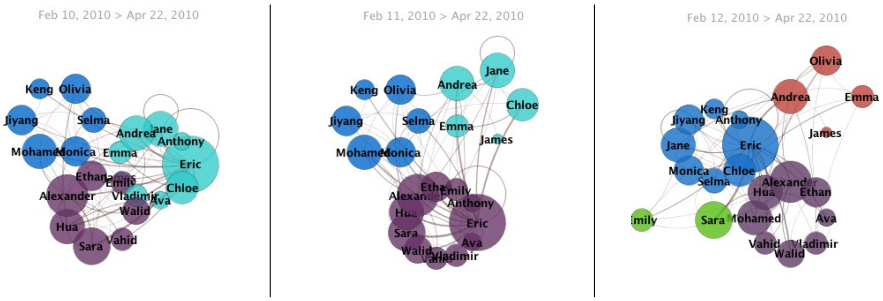


Figure 4: Monitoring changes in the student collaborative groups over time. One could observe, from left to right, how the Cyan group first looses some of its members and then splits in the final time stamp.

CONCLUSION

I believe that interconnected data analysis holds the key to understanding our world, which is becoming more connected every day. This analysis provides the much needed tools to study today's phenomena, e.g., online societies; and enables us to address the world's emerging problems, e.g., online human trafficking. My approach for interconnected data analysis brings network science, machine learning, and data mining together, leads to advances in these fields, and fills the gap in between. I am motivated by the social impacts of this research, as well as its prospect to advance the scientific methodologies we use to study our world.

REFERENCES

[1] A. ABNAR, M. TAKAFFOLI, **R. RABBANY**, AND O. R. ZAÏANE, SSRM: STRUCTURAL SOCIAL ROLE MINING FOR DYNAMIC SOCIAL NETWORKS, Social Network Analysis and Mining (SNAM), Springer, 5 (2015), pp. 1–18.

[2] D. ESWARAN, **R. RABBANY**, A. W. DUBRAWSKI, AND C. FALOUTSOS, SOCIAL-AFFILIATION NETWORKS: PATTERNS AND THE SOAR MODEL, submitted to European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), 2018.

[3] J. FAGNAN, **R. RABBANY**, M. TAKAFFOLI, E. VERBEEK, AND O. R. ZAÏANE, COMMUNITY DYNAMICS: EVENT AND ROLE ANALYSIS IN SOCIAL NETWORK ANALYSIS, in Advanced Data Mining and Applications (ADMA), Springer, 2014, pp. 85–97.

[4] **R. RABBANY**, COMMUNITY MINING AND ITS APPLICATIONS IN EDUCATIONAL ENVIRONMENTS, Master's thesis, University of Alberta, 2010.

[5] **R. RABBANY**, MODULAR STRUCTURE OF COMPLEX NETWORKS, PhD thesis, University of Alberta, 2016.

[6] **R. RABBANY**, D. BAYANI, AND A. W. DUBRAWSKI, ACTIVE LINK INFERENCE FOR CASE BUILDING INVESTIGATION, in NIPS workshop on Advances in Modeling and Learning Interactions from Complex Data (AMLICD), 2017.

[7] **R. RABBANY**, D. BAYANI, AND A. W. DUBRAWSKI, ACTIVE SEARCH OF CONNECTIONS FOR CASE BUILDING AND COMBATING HUMAN TRAFFICKING, in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2018.

[8] **R. RABBANY**, J. CHEN, AND O. R. ZAÏANE, TOP LEADERS COMMUNITY DETECTION APPROACH IN INFORMATION NETWORKS, in KDD Workshop on Social Network Mining and Analysis (SNA-KDD) , 2010.

[9] **R. RABBANY**, S. ELATIA, M. TAKAFFOLI, AND O. R. ZAÏANE, COLLABORATIVE LEARNING OF STUDENTS IN ONLINE DISCUSSION FORUMS: A SOCIAL NETWORK ANALYSIS PERSPECTIVE, Educational Data Mining: Applications and Trends, A. Peña-Ayala, ed., Springer, 2014, pp. 441–466.

[10] **R. RABBANY**, D. ESWARAN, A. W. DUBRAWSKI, AND C. FALOUTSOS, BEYOND ASSORTATIVITY: PROCLIVITY INDEX FOR ATTRIBUTED NETWORKS (PRONE), in the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2017, p. 225–237.

[11] **R. RABBANY**, D. ESWARAN, A. W. DUBRAWSKI, AND C. FALOUTSOS, PROCLIVITY PATTERNS IN ATTRIBUTED GRAPHS, in SIAM Workshop on Network Science (NS17), 2017.

[12] **R. RABBANY**, M. TAKAFFOLI, J. FAGNAN, O. R. ZAÏANE, AND R. CAMPELLO, RELATIVE VALIDITY CRITERIA FOR COMMUNITY MINING ALGORITHMS, in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Aug 2012, pp. 258–265.

[13] **R. RABBANY**, M. TAKAFFOLI, J. FAGNAN, O. R. ZAÏANE, AND R. CAMPELLO, COMMUNITIES VALIDITY: METHODICAL EVALUATION OF COMMUNITY MINING ALGORITHMS, Social Network Analysis and Mining (SNAM), Springer, 3 (2013), pp. 1039–1062.

[14] **R. RABBANY**, M. TAKAFFOLI, J. FAGNAN, O. R. ZAÏANE, AND R. CAMPELLO, RELATIVE VALIDITY CRITERIA FOR COMMUNITY MINING ALGORITHMS, Encyclopedia of Social Network Analysis and Mining (ESNAM), Springer, 2014, pp. 1562–1576.

[15] **R. RABBANY**, M. TAKAFFOLI, AND O. R. ZAÏANE, ANALYZING PARTICIPATION OF STUDENTS IN ONLINE COURSES USING SOCIAL NETWORK ANALYSIS TECHNIQUES, in Educational Data Mining (EDM), 2011, pp. 21–30.

[16] **R. RABBANY**, M. TAKAFFOLI, AND O. R. ZAÏANE, SOCIAL NETWORK ANALYSIS AND MINING TO SUPPORT THE ASSESSMENT OF ON-LINE STUDENTnipsW17 PARTICIPATION, ACM SIGKDD Explorations Newsletter, 13 (2011), pp. 20–29.

[17] **R. RABBANY** AND O. R. ZAÏANE, A DIFFUSION OF INNOVATION-BASED CLOSENESS MEASURE FOR NETWORK ASSOCIATIONs, in IEEE International Conference on Data Mining (ICDM) Workshops, 2011, pp. 381–388.

[18] **R. RABBANY** AND O. R. ZAÏANE, EVALUATION OF COMMUNITY MINING ALGORITHMS IN THE PRESENCE OF ATTRIBUTES, Trends and Applications in Knowledge Discovery and Data Mining, Springer, 2015, pp. 152–163.

[19] **R. RABBANY** AND O. R. ZAÏANE, GENERALIZATION OF CLUSTERING AGREEMENTS AND DISTANCES FOR OVERLAPPING CLUSTERS AND NETWORK COMMUNITIES, Data Mining and Knowledge Discovery (DAMI), Springer, 29 (2015), pp. 1458–1485.

[20] **R. RABBANY** AND O. R. ZAÏANE, A GENERAL CLUSTERING AGREEMENT INDEX: FOR COMPARING DISJOINT AND OVERLAPPING CLUSTERS, in Thirty-First AAAI Conference on Artificial Intelligence (AAAI), 2017, pp. 2492–2498.

[21] S. RAVANBAKHSH, **R. RABBANY**, AND R. GREINER, AUGMENTATIVE MESSAGE PASSING FOR TRAVELING SALESMAN PROBLEM AND GRAPH PARTITIONING, in Advances in Neural Information Processing Systems (NIPS), 2014, pp. 289–297.

[22] C. LARGERON, P.-N. MOUGEL, **R. RABBANY**, AND O. R. ZAÏANE, GENERATING ATTRIBUTED NETWORKS WITH COMMUNITIES, Public Library of Science (PLoS ONE), 10 (2015), p. e0122777.

[23] M. TAKAFFOLI, **R. RABBANY**, AND O. R. ZAÏANE, INCREMENTAL LOCAL COMMUNITY IDENTIFICATION IN DYNAMIC SOCIAL NETWORKS, in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2013, pp. 90–94.

[24] M. TAKAFFOLI, **R. RABBANY**, AND O. R. ZAÏANE, COMMUNITY EVOLUTION PREDICTION IN DYNAMIC SOCIAL NETWORKS, in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2014, pp. 9–16.