Research Statement

Cuneyt Gurcan Akcora

University of Texas at Dallas

My principal research interests lie in statistics, machine learning, graph mining, data science, and data-driven statistical inference on complex networks. Nowadays, our daily lives are shaped around how we use various networks, ranging from online social media to dating websites to cryptocurrency and Blockchain networks. In my research I aim to create novel machine learning and statistical algorithms and tools to model and analyze these networks.

A recurring theme in my research is that of modeling *local* interactions on complex networks. Interactions create nonlinear temporal and spatial structures that shape network communities and the overall network topology. In turn, studying how and why nodes interact is a critical step toward understanding the functionality and organization of complex networks, and has high utility in a variety of applications. For example, my research answers diverse questions such as *How are friendships formed on Facebook?*, *How do opinions change on Twitter?*, and *How do transfers among address nodes impact the Bitcoin price?*.

Throughout my research, I focused on two main research objectives. My first goal is to solve real-world problems on very large-scale datasets [17]; this includes recommending friends, identifying abnormal user behavior and predicting Bitcoin price. The second objective is to understand and explain the underlying network mechanisms at play; to learn how people view their social experience and interact with others, and to aggregate seemingly local effects to create global knowledge. My success in these two goals have resulted in works that have already received hundreds of peer citations.

Below I discuss main directions in my interdisciplinary research agenda: analysis of large complex networks by studying local graph structures (Section 1), development of efficient models to explain user behavior on social networks (Section 2), and measurement of how users participate in the learning and opinion forming on a network (Section 3). I show that these research directions can build novel applications, lead to more accurate predictions than currently available forecasting tools, and allow to enhance our understanding of how complex networks will evolve.

1 Graph Mining and Inference on Complex Networks

Technological advances have diversified and revolutionized user data. Mobile phones save our location, and search engines track what we search online. From cycling routes to shopping lists, there exist applications to store any kind of digital footprint we create. With users sharing this wealth of information online, recent years have seen a proliferation of complex networks.

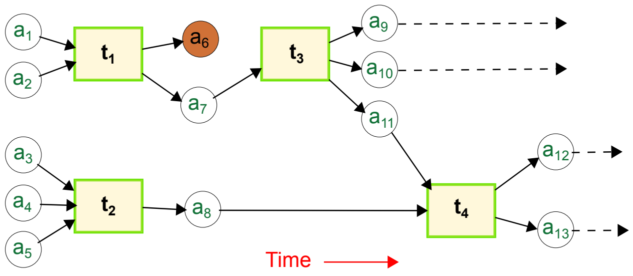

Figure 1: Shaped with decisions of humans, machines and automated scripts, the heterogeneous Blockchain networks are a prime example of complex networks. Our work predicts Bitcoin price with local structures on these networks [10].

In my research, I employed complex network analysis on social [11], scholarly [4], location [15], web-knowledge [1] and Blockchain networks [14]. In these articles, my goal is to assess and model network local higher-order structures, such as friendship motifs [11] on social networks and chainlets [10] in the Bitcoin network.

The rationale behind my interest to delve into the analysis of these local network topologies is the following. Local higher-order structures are an indispensable tool for analysis of network organization beyond the trivial scale of individual vertices and edges. The role of small subgraphs, also called network *motifs* or *graphlets*, has been first discussed in conjunction with the assessment of stability and robustness of biological networks, and later have been studied in a variety of contexts. However, local higher-order topologies are much less investigated in online social networks and remain virtually unexplored in financial systems, particularly, in Blockchain.

We proposed the first local structures, called chainlets, on the Bitcoin network. A chainlet is the smallest subgraph on the Bitcoin network, and its type depends on the inputs and outputs of a transaction. My results on Bitcoin show the utility of local chainlets; looking at the full historical data since 2008, I assessed the role of 400 distinct types of chainlets on Bitcoin price. Our findings indicate that certain types of chainlets exhibit the highest predictive influence on Bitcoin price and investment risk. Moreover, my work shows that depending on the nature of the analysis, such as anomaly detection or price prediction, the Bitcoin network should make use of specific chainlets only. I believe that these findings will strongly influence how other researchers will use the Bitcoin network.

The efficiency of motifs and chainlets emphasizes the importance of intrinsic geometry in graph mining. However, analysis of data shape becomes much more challenging for weighted networks, as methods for inference on weighted motifs are still quite limited. As an alternative, I now explore application of topological data analysis on weighted dynamic networks. The core idea to this challenging problem is to embed a network into a geometric space and then to analyze the so-called *simplical complexes*, or a set of elementary objects such as points, line segments, triangles, tetrahedra and their higher-dimensional analogs. Persistent homology, or analysis of properties of progressively finer simplical complexes, then unveils some critical properties behind functionality of complex networks at multi-scale levels, which are otherwise largely unaccessible with conventional analytical approaches.

Furthermore, a major aspect of graph mining on complex networks deals with validating the results, or uncertainty quantification in the reported findings. In some of my studies, I have used such statistical or information theoretic measures, as Fleiss' Kappa [8] and the information gain ratio [5]. However, in some cases inference on social networks requires human validation. For example, in [8] I created an anomaly detection framework on the Twitter network by considering both connections and tweets of users. The framework detects anomalous user behavior, but the inferred results do not have a ground truth to compare against. To overcome this problem, we used an Amazon Mechanical Turk validation scheme, where humans were presented with a validation interface to rate our inferences.

Although useful, the Mechanical Turk experiments were costly, hence it cannot be done in a large scale. In a recent work, we addressed the problem of quantifying the confidence in network analysis [15] with bootstrapped sampling of features. This work shed light on key questions in social network analysis such as: given a limited availability of social network data, how much data should be queried from the network, and which node features can be learned reliably? More importantly, how can we evaluate the uncertainty of our estimators? To address these challenges, we proposed a novel bootstrap method for uncertainty analysis of node features in social network mining, derived its asymptotic properties, and demonstrated its effectiveness with extensive experiments in large real life networks. We are now expanding methods of bootstrap inference on graphs to dynamic anomaly detection and uncertainty quantification in estimation of higher-order structures.

2 Behavior Modeling on Complex Networks

More than using a purely graph oriented approach, in my research works I aimed at using data from nodes and edges in explaining how networks function and change. Although this data is unstructured, noisy or plain wrong in many cases [4], using it along with the graph data facilitates learning the rationale behind discovered results. For example, we used unstructured user data from Facebook profiles, and devised two measures to predict friendships [5, 4].

Having an account on an Online Social Network (OSN) opens a path to opportunities but it also brings about certain risks [6]; social network users can be bullied, their pictures can be stolen or their status posts can reach unwanted audiences. Even when profiles do not list any information, social graphs can be analyzed to infer personal information. Despite all these, social networks have hundreds of millions of users, because users think that positives outweigh negatives and they maintain their online presence. However, this user attitude is not totally care-free; individual cases of privacy breaches and their consequences have been widely discussed in social media and privacy risks have only grown with time as social networks have grown in size exponentially [12].

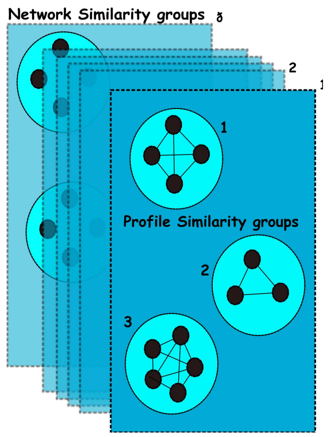

Figure 2: Our widely adopted similarity approach compares users in terms of shared friendships and attributes [5].

I adopted an active learning approach for risk estimation, where a user's risk attitude is learned from few required interactions. With our learning algorithm, once the classifier is built with the training data, the system can predict risk labels of all those strangers that were not included in the training data without any user intervention.

The discussed risk estimation process has been developed into a Facebook application and tested on real Facebook data. Our experiments show that with limited user involvement, we can get accurate risk estimation. In particular, our experiments show that we are able to correctly predict risk labels with 83.38% accuracy.

As we could explain similarities and differences of users with our measures [7, 13], this line of research directly mapped individual user behavior to the widely used theories (e.g., Homophily) from Sociology. As a further benefit, this allowed us to see the network from the users' perspective, and quantify risks of friendships [6]. Both the approach to explain user behavior, and the two similarity measures have been widely adopted by other researchers.

My research in complex networks have used network data to predict how nodes and edges will change the network [3]. In addition to observing the network in this fashion, I am involved into project entitled "Human Pathways" that studies and tracks capabilities, decisions and career choices of humans in large-scale datasets from social networks, news media and surveys. The pilot algorithms of Human Pathways focus on developing optimal retention strategies for minorities and underrepresented groups as a part of NSF INCLUDES. My goal in this project is to impact lives of humans positively, by automatically recommending them entire pathways to be taken in their professional lives.

3 Participation and Opinion Formation on Networks

My experience in software development has enabled me to create novel research applications. Some examples of my software projects are the anomaly detection framework that employs Amazon Mechanical Turk experiments [8], the browser extension and Flash based validation website in [5] and the web-knowledge graph rule detection framework in [1]. Specifically, my work in the Upinion project [16, 2] shows how an application framework can be used in a novel research area.

I developed the Upinion project to show the benefits of active data mining where we can tap into the vast social network posts that are publicly available. The project employed a backend that was integrated with the Twitter Java [9] API and Google Search API for two applications.

In the first application [2], Upinion tracked public opinion about a given topic (e.g., Obama's presidency in 2009) by simply monitoring Twitter posts and mapping the posts into six different emotion classes such as sadness and anger. The project accepted topic subscriptions, and had an interactive interface to show user posts from each emotion class about a given topic. I employed vector based similarities to detect periods where public opinion was changing. Furthermore, through a Google Search API I listed all possible events that led to the opinion change. Rather than conducting costly opinion polls, I showed that public opinion about a topic can be tracked on Twitter in an automated fashion. This work has proved very influential; its results have been used as a comparison method in tens of research works afterwards.

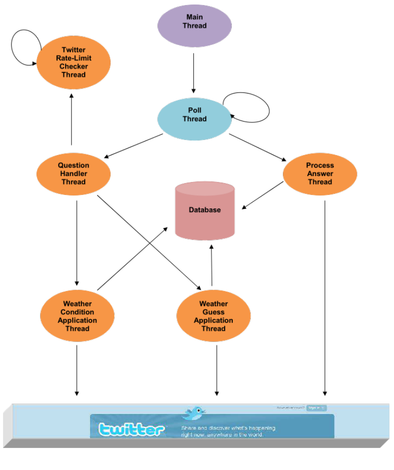
Main Thread
Twitter Rate-Limit Checker Thread
Poll Thread
Question Handler Thread
Process Answer Thread
Database
Weather Condition Application Thread
Weather Guess Application Thread
twitter

Figure 3: Our crowd-sourced sending application on Twitter was the first attempt to learn from social network users [16].

In the second application [16], we proposed the first open infrastructure on Twitter to learn from users and paved the way for ubiquitous crowd-sourcing and collaboration applications. The crowd-sourcing system architecture over Twitter was used in two case studies: weather radar and noise mapping. Even without an incentive structure, Twitter users volunteered to participate in the crowd-sourcing experiments (with around 15% reply rates) and the latency of the replies were low. Up to 50% of replies arrived in 30 minutes and 80% of replies arrived in 2 hours.

The success of the Upinion project showed the feasibility of learning about the real world from user generated data. Furthermore, in my early research career its success instilled me with the motivation to further integrate my machine learning and software development skills with more advanced statistical methodology.

References

[1] Z. Abedjan, C. G. Akcora, M. Ouzzani, P. Papotti, and M. Stonebraker. Temporal rules discovery for web data cleaning. *Proceedings of the VLDB Endowment*, 9(4):336–347, 2015.

[2] C. G. Akcora, M. A. Bayir, M. Demirbas, and H. Ferhatosmanoglu. Identifying breakpoints in public opinion. In *Proceedings of the first workshop on social media analytics*, pages 62–66. ACM, 2010.

[3] C. G. Akcora, B. Carminati, and E. Ferrari. Building virtual communities on top of online social networks. In *European Conference on Information Management and Evaluation*, page 19. Academic Conferences International Limited, 2011.

[4] C. G. Akcora, B. Carminati, and E. Ferrari. Network and profile based measures for user similarities on social networks. In *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*, pages 292–298. IEEE, 2011.

[5] C. G. Akcora, B. Carminati, and E. Ferrari. Privacy in social networks: How risky is your social graph? In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 9–19. IEEE, 2012.

[6] C. G. Akcora, B. Carminati, and E. Ferrari. Risks of friendships on social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 810–815. IEEE, 2012.

[7] C. G. Akcora, B. Carminati, and E. Ferrari. User similarities on social networks. *Social Network Analysis and Mining*, 3:1–21, 2013.

[8] C. G. Akcora, B. Carminati, E. Ferrari, and M. Kantarcioglu. Detecting anomalies in social network data consumption. *Social Network Analysis and Mining*, 4(1):231, 2014.

[9] C. G. Akcora and M. Demirbas. Twitter: Roots, influence, applications. *Department of Computing Science and Engineering, SUNY Buffalo, January*, 2010.

[10] C. G. Akcora, A. K. Dey, Y. R. Gel, and M. Kantarcioglu. Forecasting bitcoin price with graph chainlets. *The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2018.

[11] C. G. Akcora and E. Ferrari. Discovering trust patterns in ego networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 224–229. IEEE, 2014.

[12] C. G. Akcora and E. Ferrari. Graphical user interfaces for privacy settings. In *Encyclopedia of Social Network Analysis and Mining*. Springer, 2014.

[13] C. G. Akcora and E. Ferrari. Similarity metrics on social networks. In *Encyclopedia of Social Network Analysis and Mining*. Springer, 2014.

[14] C. G. Akcora, Y. R. Gel, and M. Kantarcioglu. Blockchain: A graph primer. *arXiv preprint arXiv:1708.08749*, 2017.

[15] C. G. Akcora, Y. R. Gel, and M. Kantarcioglu. Quantifying uncertainty in node feature analysis of large social networks. *Under submission*, 2018.

[16] M. Demirbas, M. A. Bayir, C. G. Akcora, Y. S. Yilmaz, and H. Ferhatosmanoglu. Crowd-sourced sensing and collaboration using twitter. In *World of Wireless Mobile and Multimedia Networks (WoWMoM), 2010 IEEE International Symposium on a*, pages 1–9. IEEE, 2010.

[17] W. Lucia, C. G. Akcora, and E. Ferrari. Multi-dimensional conversation analysis across online social networks. In *Social Computing and its applications (SCA), 2013 Third International Conference on*, pages 331–336. IEEE, 2013.