Abstracts of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Best Papers

Influence based Analysis of Community Consistency in Dynamic Networks

Xiaowei Jia, Xiaoyi Li, Nan Du, Yuan Zhang, Vishrawas Gopalakrishnan, Guangxu Xun and Aidong Zhang

The development of Internet and social networks has provided more emerging network data which facilitates the dynamic network analysis. In this paper, we propose anew method to measure coherence strength, also referred to as community consistency, of a community under dynamic settings. In order to better interpret the influence of evolving community structure on community consistency, we model the problem as one of influence propagation processes having a causal relation with the community consistency. To this effect a generative model is proposed to combine the influence propagation and the network topological structure at each time stamp. Our comprehensive experiments on both synthetic and real-world datasets demonstrate the superiority of the proposed frameworkin estimating the community consistency.

Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla and Niloy Ganguly

Most of the online news media outlets rely heavily on the revenues generated from the clicks made by their readers, and due to the presence of numerous such outlets, they need to compete with each other for reader attention. To attract the readers to click on an article and subsequently visit the media site, the outlets often come up with catchy headlines accompanying the article links, which lure the readers to click on the link. Such headlines are known as Clickbaits. While these baits may trick the readers into clicking, in the long run, clickbaits usually don't live up to the expectation of the readers, and leave them disappointed. In this work, we attempt to automatically detect clickbaits and then build a browser extension which warns the readers of different media sites about the possibility of being baited by such headlines. The extension also offers each reader an option to block clickbaits she doesn't want to see. Then, using such reader choices, the extension automatically blocks similar clickbaits during her future visits. We run extensive offline and online experiments across multiple media sites and find that the proposed clickbait detection and the personalized blocking approaches perform very well achieving % accuracy in detecting and % accuracy in blocking clickbaits.

A 1: Graphs

Streaming METIS Partitioning

Ghizlane Echbarthi and Hamamache Kheddouci

The proliferation in size of actual graph datasets impels the use of distributed graph processing frameworks which in turn, should consider a good partitioning of the graph dataset in order to see their performances enhanced. In this paper, we focus on a well known heuristic for graph partitioning named METIS, an offline method giving high quality partitions but unsuitable for processing large graphs due to the offline setting. A recently proposed alternative is the streaming partitioning heuristics aiming to alleviate the computational

resources constraints when dealing with large graphs. In considering this matter, we propose a new partitioning method that benefits from the accuracy of METIS and the lightness of the streaming setting. This work introduces the Streaming METIS Partitioning method (SMP) which is an online counterpart of METIS, a fast and well known multilevel heuristic for graph partitioning. We show in a complexity analysis that SMP has a lower time complexity compared to METIS, which is confirmed by conducted experiments. Moreover, we show that SMP yields competitive results to its offline counterpart METIS, especially when it is run on a Depth First Search streaming order. Also, when compared to other online competitors, SMP is the best performing heuristic giving partitions with minimal edge cut.

New Stopping Criteria For Spectral Partitioning

James Fairbanks, Anita Zakrzewska and David A. Bader

Spectral partitioning (clustering) algorithms use eigenvectors to solve network analysis problems. The relationship between numerical accuracy and network mining quality is insufficiently understood. We show that analyzing numerical accuracy and network mining quality together leads to an algorithmic improvement. Specifically, we study spectral partitioning using sweep cuts of approximate eigenvectors of the normalized graph Laplacian. We introduce a novel, theoretically sound, parameter free stopping criterion for iterative eigensolvers designed for graph partitioning. On a corpus of social networks, we validate this stopping criterion by showing the number of iterations is reduced by a factor of : on average, and the conductance is increased by only a factor of : on average. Regression analysis of these results shows that the decrease in the number of iterations needed is greater for problems with a small spectral gap, thus our stopping criterion helps more on harder problems. Experiments show that alternative stopping criteria are insufficient to ensure low conductance partitioning on real world networks. While our method guarantees partitions that satisfy the Cheeger Inequality, we find that it typically beats this guarantee on real world graphs.

Local Triangle-Densest Subgraphs

Raman Samusevich, Maximilien Danisch and Mauro Sozio

Finding dense subgraphs in large graphs is a key primitive in a variety of real-world application domains, encompassing social network analysis, event detection, problems arising in biology and many others. Several recent works have studied some variants of the classical densest subgraph problem, considering alternative quality measures such as the average number of triangles in the subgraphs and their compactness. Those are desirable properties when the task is to find communities or interesting events in social networks. In our work, we capitalize on previous works and study a variant of the problem where we aim at finding subgraphs which are both compact and contain a large number of triangles. We provide a formal definition for our problem, while developing efficient algorithms with strong theoretical guarantees. Our experimental evaluation on large real-world networks shows the effectiveness of our approach.

Tradeoffs between Density and Size in Extracting Dense Subgraphs: A Unified Framework

Zhefeng Wang, Lingyang Chu, Jian Pei, Abdullah Al-Barakati and Enhong Chen

Extracting dense subgraphs is an important step in many graph related applications. There is a challenging struggle in exploring the tradeoffs between density and size in subgraphs extracted. More often than not, different methods aim at different specific tradeoffs between the two factors. To the best of our knowledge, no existing method can allow a user to explore the full spectrum of the tradeoffs using a single parameter. In this paper, we investigate this problem systematically. First, since the existing studies cannot find highly compact dense subgraphs, we formulate the problem of finding very dense but relatively small subgraphs. Second, we connect our problem with the existing methods and propose a unified framework that can explore the tradeoffs

between density and size of dense subgraphs extracted using a hyper-parameter.We give theoretical upper and lower bounds on the hyper-parameter so that the range where the unified framework can produce non-trivial subgraphs is determined. Third, we develop an efficient quadratic programming method for the unified framework, which is a generalization and extension to the existing methods. We show that optimizing the unified framework is essentially a relaxation of the maximization of a family of density functions. Last, we report a systematic empirical study to verify our findings.

B 1: Communities

Non-Sharing Communities? An Empirical Study of Community Detection for Access Control Decisions

Gaurav Misra, Jose M. Such and Hamed Balogun

Social media users often find it difficult to make appropriate access control decisions which govern how they share their information with a potentially large audience on these platforms. Community detection algorithms have been previously put forth as a solution which can help users by automatically partitioning their friend network. These partitions can then be used by the user as a basis for making access control decisions. Previous works which leverage communities for enhancing access control mechanisms assume that members of the same community will have the same access to a user's content, but whether or to what extent this assumption is correct is a lingering question. In this paper, we empirically evaluate a goodness of fit between the communities created by implementing community detection algorithms on the friend networks of users and the access control decisions made by them during a user study. We also analyze whether personal characteristics of the users or the nature of the content play a role in the performance of the algorithms. The results indicate that community detection algorithms may be useful for creating default access control policies for users who exhibit a relatively more static access control behaviour. For users showing great variation in their access control decisions across the board (both in terms of number and actual members), we found that community detection algorithms performed poorly.

Functional Cluster Extraction from Large Spatial Networks

Takayasu Fushimi, Kazumi Saito, Tetsuo Ikeda and Kazuhiro Kazama

We address a problem of extracting functionally similar regions in urban streets regarded as spatial networks. Such characteristics of regions will play important roles for developing and planning city promotion, travel tours and so on, as well as understanding and improving the usage of urban streets. In order to analyze such functionally similar regions, we propose an acceleration method of the FCE (functional cluster extraction) algorithm equipped with the lazy evaluation and pivot pruning techniques, which enables to efficiently deal with several largescale networks. In our experiments using urban streets of six cities, we show that our proposed method achieved a reasonably high acceleration performance. Then, we show that functional cluster produced by our method are useful for understanding the properties of areas in a series of visualization results.

Network Completion via Joint Node Clustering and Similarity Learning

Dimitrios Rafailidis and Fabio Crestani

In this study, we investigate the problem of network completion by considering the similarities between the node attributes. Given a sample of observed nodes with their incident edges, how can we efficiently reconstruct the network by completing the missing edges of unobserved nodes? Apart from the missing edges, in real settings the node attributes may be partially missing, as well as they may introduce noise when completing the network. We propose a network completion method based on joint clustering and similarity learning. The proposed approach differs from competitive strategies, which consider attribute-based

similarities at the node-level. First we generate clusters based on the node attributes, thus reducing the noise and the sparsity in the case that the attributes may be missing. We design a joint objective function to jointly factorize the adjacency matrix of the observed edges with the clusterbased similarities of the node attributes, while at the same time the clusters are adapted, accordingly. In addition, we propose an optimization algorithm to solve the network completion problem via alternating minimization. Our experiments on two real world social networks from Facebook and Google+ show that the proposed approach achieves high completion accuracy, compared to other state-of-the-art methods.

Sensitivity and Reliability in Incomplete Networks: Centrality Metrics to Community Scoring Functions

Soumya Sarkar, Suhansanu Kumar, Sanjukta Bhowmick and Animesh Mukherjee

In this paper we evaluate the effect of noise ons coring and centrality-based parameters with respect to two different aspects of network analysis: (i) sensitivity, that is how the parameter value changes as edges are removed and (ii) reliability in the context of message spreading, that is how the time taken to broadcast a message changes as edges are removed. Our experiments on synthetic and real-world networks and three different noise models demonstrate that for both the aspects over all networks and all noise models, permanence qualifies as the most effective metric. For the sensitivity experiments closeness centrality is a close second. For the message spreading experiments, closeness and betweenness centrality based initiator selection closely competes with permanence. This is because permanence has a dual characteristic where the cumulative permanence over all vertices is sensitive to noise but the ids of the top-rank vertices, which are used to find seeds during message spreading remain relatively stable under noise.

Ensemble-Based Algorithms to Detect Disjoint and Overlapping Communities in Networks

Tanmoy Chakraborty, Noseong Park and V.S. Subrahmanian

Given a set AL of community detection algorithms and a graph G as inputs, we propose two ensemble methods EnDisCo and MeDOC that (respectively) identify disjoint and overlapping communities in G. EnDisCo transforms a graph into a latent feature space by leveraging multiple base solutions and discovers disjoint community structure. MeDOC groups similar base communities into a meta-community and detects both disjoint and overlapping community structures. Experiments are conducted at different scales on both synthetically generated networks as well as on several real-world networks for which the underlying ground-truth community structure is available. Our extensive experiments show that both algorithms outperform state-of-the-art non-ensemble algorithms by a significant margin. Moreover, we compare EnDisCo and MeDOC with a recent ensemble method for disjoint community detection and show that our approaches achieve superior performance. To the best of our knowledge, MeDOC is the first ensemble approach for overlapping community detection.

C 1: Politics, Unrest

Community Detection in Political Twitter Networks using Nonnegative Matrix Factorization Methods

Mert Ozer, Nyunsu Kim and Hasan Davulcu

Community detection is a fundamental task in social network analysis. In this paper, first we develop an endorsement filtered user connectivity network by utilizing Heider's structural balance theory and certain Twitter triad patterns. Next, we develop three Nonnegative Matrix Factorization frameworks to investigate the contributions of different types of user connectivity and content information in community detection. We show that user content and endorsement filtered connectivity information are complementary to each other in clustering politically motivated users into pure political communities. Word usage is the strongest indicator of

users' political orientation among all content categories. Incorporating user-word matrix and word similarity regularizer provides the missing link in connectivity only methods which suffer from detection of artificially large number of clusters for Twitter networks.

On Predicting Social Unrest Using Social Media

Rostyslav Korolov, Di Lu, Jingjing Wang, Guangyu Zhou, Claire Bonial, Clare Voss, Lance Kaplan, William Wallace, Jiawei Han and Heng Ji

We study the possibility of predicting a social protest (planned, or unplanned) based on social media messaging. We consider the process called mobilization, described in the literature as the precursor of participation. Mobilization includes four stages: being sympathetic to the cause, being aware of the movement, motivation to take part and ability to participate. We suggest that expressions of mobilization in communications of individuals may be used to predict the approaching protest. We have utilized several Natural Language Processing techniques to create a methodology to identify mobilization in social media communication. Results of experimentation with Twitter data collected before and during the Baltimore events and the information on actual protests taken from news media show a correlation over time between volume of Twitter communications related to mobilization and occurrences of protest at certain geographical locations. We conclude with discussion of possible theoretical explanations and practical applications of these results.

Characterizing Communal Microblogs during Disaster Events

Koustav Rudra, Ashish Sharma, Niloy Ganguly and Saptarshi Ghosh

Millions of microblogs are posted during disasters, which include not only information about the present situation, but also the emotions / opinions of the masses. While most of the prior research has been on extracting situational information, this work focuses on a particular type of non-situational tweets – communal tweets, i.e., abusive posts targeting specific religious / racial groups. We characterize the communal tweets posted during five recent disaster events, and the users who posted such tweets. We find that communal tweets are posted not only by common users, but also by many popular users (having tens of thousands of followers), most of whom are related to the media and politics. As a result, communal tweets get much higher exposure (retweets) than non-communal tweets. Further, users posting communal tweets form strong connected groups in the social network. Considering the potentially adverse effects of communal tweets during disasters, we also indicate a way to counter such tweets, by utilizing anti-communal tweets posted by some users during such events.

Investigating the complete corpus of Referendum and Elections tweets

Despoina Antonakaki, Dimitris Spiliotopoulos, Christos V. Samaras, Sotiris Ioannidis and Paraskevi Fragopoulou

Today, a considerable proportion of the public political discourse that proceeds nationwide elections is happening through Online Social Networks. Through analyzing this content, we can discover the major themes that prevailed during the discussion, investigate the temporal variation of positive and negative sentiment and examine the semantic proximity of these themes. According to existing studies, the results of similar tasks are heavily dependent on the quality and completeness of dictionaries for linguistic preprocessing, entity discovery and sentiment analysis. Additionally, noise reduction is achieved with methods for sarcasm detection and correction. Here we report on the application of these methods on the complete corpus of tweets regarding two local electoral events of worldwide impact: the Greek referendum of and the subsequent legislative elections. To this end, we compiled novel dictionaries for sentiment and entity detection for the Greek language tailored to these events. We subsequently performed volume analysis, sentiment

analysis and sarcasm correction. Results showed that there was a strong anti-austerity sentiment accompanied with a critical view on European and Greek political actions.

Understanding Citizen Reactions and Ebola-Related Information Propagation on Social Media

Thanh Tran and Kyumin Lee

In severe outbreaks such as Ebola, bird flu and SARS, people share news, and their thoughts and responses regarding the outbreaks on social media. Understanding how people perceive the severe outbreaks, what their responses are, and what factors affect these responses become important. In this paper, we conduct a comprehensive study of understanding and mining the spread of Ebola-related information on social media. In particular, we (i) conduct a large-scale data-driven analysis of geotagged social media messages to understand citizen reactions regarding Ebola; (ii) build information propagation models which measure locality of information; and (iii) analyze spatial, temporal and social properties of Ebola-related information. Our work provides new insights into Ebola outbreak by understanding citizen reactions and topic-based information propagation, as well as providing a foundation for analysis and response of future public health crises.

A 2: Sampling and Streaming

Estimating Exponential Random Graph Models using Sampled Network Data via Graphon

Ran He and Tian Zheng

Analysis of large networks is of interest to many disciplines. Full network data are often hard to collect, storage and analyze. In particular, in many social science surveys, ego nomination techniques have been used to collect the egocentric networks of the randomly sampled survey respondents. In this paper, we propose a sample-GLMLE method that fits exponential random graph models (ERGM) to such sampled egocentric networks. It is an extension of a previous graph-limit based maximum likelihood estimation (GLMLE) method for full network that uses graphon to bridge the estimation of ERGM using observed network data. In this paper, we provide solutions to computational issues that are unique to sampled network data and evaluate the proposed method using simulations. We also apply sample-GLMLE to the public-use set of the National Longitudinal Study of Adolescent Health (AddHealth) study.

Rank Degree: An Efficient Algorithm for Graph Sampling

Elli Voudigari, Nikos Salamanos, Theodore Papageorgiou and Emmanuel Yannakoudakis

The study of a large real world network in terms of graph sample representation constitutes a very powerful and useful tool in several domains of network analysis. This is the motivation that has led the work of this paper towards the development of a new graph sampling algorithm. Previous research in this area proposed simple processes such as the classic Random Walk algorithm, Random node and Random edge sampling and has evolved during the last decade to more advanced graph exploration approaches such as Forest Fire and Frontier sampling. In this paper, we propose a new graph sampling method based on edge selection. In addition, we crawled Facebook collecting a large dataset consisting of million users and million users' relations, which we have also used to evaluate our sampling algorithm. The experimental evaluation on several datasets proves that our approach preserves several properties of the initial graphs, leading to representative samples and outperforms all the other approaches.

Query-Friendly Compression of Graph Streams

Arijit Khan and Charu Aggarwal

We study the problem of synopsis construction of massive graph streams arriving in real-time.Many graphs such as those formed by the activity on social networks, communication networks, and telephone networks are defined dynamically as rapid edge streams on a massive domain of nodes. In these rapid and massive graph streams, it is often not possible to estimate the frequency of individual items (e.g., edges, nodes) with complete accuracy. Nevertheless, sketch-based stream summaries such as Count-Min can preserve frequency information of highfrequency items with a reasonable accuracy. However, these sketch summaries lose the underlying graph structure unless one keeps information about start and end nodes of all edges, which is prohibitively expensive. For example, the existing methods can identify the high-frequency nodes and edges, but they are unable to answer more complex structural queries such as reachability defined by high-frequency edges. To this end, we design a -dimensional sketch, gMatrix that summarizes massive graph streams in real-time, while also retaining information about the structural behavior of the underlying graph dataset. We demonstrate how gMatrix, coupled with a onetime reverse hash mapping, is able to estimate important structural properties, e.g., reachability over high frequency edges in an online manner and with theoretical performance guarantees. Our experimental results using large-scale graph streams attest that gMatrix is capable of answering both frequency-based and structural queries with high accuracy and efficiency.

Classification in Dynamic Streaming Networks

Yibo Yao and Lawrence Holder

Traditional network classification techniques will become computationally intractable when applied on a network which is presented in a streaming fashion with continuous updates. In this paper, we examine the problem of classification in dynamic streaming networks, or graphs. Two scenarios have been considered: the graph transaction scenario and the one large graph scenario. We propose a unified framework consisting of three components: a subgraph extraction method, an online version of an existing graph kernel, and two kernel-based incremental learners. We demonstrate the advantages of our framework via empirical evaluations on several real-world network datasets.

B 2: Social Networks

Social Network Dominance based on Analysis of Asymmetry

Yuemeng Li, Xintao Wu and Song Yang

We focus on analysis of dominance, power, influence—that by definition asymmetric—between pairs of individuals in social networks. We conduct dominance analysis based on the canonical analysis of asymmetry that decomposes a square asymmetric matrix into two parts, a symmetric one and a skewsymmetric one, and then applies the singular value decomposition (SVD) on the skew-symmetric part. Each individual node can be projected as one -dimensional point based on its row values at each pair of successive singular vectors. The asymmetric relationship between two individuals can then be captured by areas of triangles formed from the two points and the origin in each -dimensional space. We quantify node dominance (submissive) score based on the relative position of the node's coordinate from coordinates of all other nodes it dominates (subdues) in the projected singular vector spaces. We conduct dominance/submissiveness analysis for several representative networks including perfect linear orderings, networks with tree structure, and networks with random graphs and examine the departures of a real social network from those representative graphs. Empirical evaluations demonstrate the effectiveness of the proposed approach.

MaxReach: Reducing Network Incompleteness through Node Probes

Sucheta Soundarajan, Tina Eliassi-Rad, Brian Gallagher and Ali Pinar

Real-world network datasets are often incomplete. Subsequently, any analysis on such networks is likely to produce skewed results. We examine the following problem: given an incomplete network, which b nodes should be probed to bring as many new nodes as possible into the observed network? For instance, consider someone who has observed a portion (say %) of the Twitter network. How should she use a limited budget to reduce the incompleteness of the network? In this work, we propose a novel algorithm, called MAXREACH, which uses a budget b to increase the number of nodes in the observed network. Our experiments, across a range of datasets and conditions, demonstrate the efficacy of MAXREACH.

Predicting Anchor Links between Heterogeneous Social Networks

Sina Sajadmanesh, Hamid R. Rabiee and Ali Khodadadi

People usually get involved in multiple social networks to enjoy new services or to fulfill their needs. Many new social networks try to attract users of other existing networks to increase the number of their users. Once a user (called source user) of a social network (called source network) joins a new social network (called target network), a new inter-network link (called anchor link) is formed between the source and target networks. In this paper, we concentrated on predicting the formation of such anchor links between heterogeneous social networks. Unlike conventional link prediction problems in which the formation of a link between two existing users within a single network is predicted, in anchor link prediction, the target user is missing and will be added to the target network once the anchor link is created. To solve this problem, we propose an effective general meta-path-based approach called Connector and Recursive Meta-Paths (CRMP). By using those two different categories of meta-paths, we model different aspects of social factors that may affect a source user to join the target network, resulting in the formation of a new anchor link. Extensive experiments on real-world heterogeneous social networks demonstrate the effectiveness of the proposed method against the recent methods.

Benchmarking Online Social Networks

Pablo Nicolas Terevinto, Miguel Perez, Josep Domenech, Jose A. Gil and Ana Pont

Although OSNs are major and growing large scale web applications, there is still a lack of workload models and tools for performance evaluation and testability studies. This fact motivates us to develop a general purpose benchmark for evaluating the main hardware and software resources associated to this kind of applications. To this end, we have developed a flexible workload model based on interactive users that, together with a complete and fully operative framework, permits to monitor system resources to perform fine grain performance and testability studies.

On the #_-Hyperbolicity in Complex Networks

Hend Alrasheed

δ-Hyperbolicity is a graph parameter that shows how close to a tree a graph is metrically. In this work, we propose a method that reduces the size of the graph to only a subset that is responsible for maximizing its δ-hyperbolicity using the local dominance relationship between vertices. Furthermore, we empirically show that the hyperbolicity of a graph can be found in a set of vertices that are in close proximity. That is, the hyperbolicity in graphs is, to some extent, a local property. Moreover, we show that this set is close to the graph's center. Our observations have crucial implications on computing the value of the δ-hyperbolicity of graphs.

Why not Scale Free? Simulating Company Ego Networks on Twitter

Yoav Achiam, Inbal Yahav and David Schwartz

This paper simulates Companies' ego networks on Twitter, meaning the companies' number and type of followers. Evident from our data, we show that followers' distribution, in our focus, is neither scale free nor random, thus common network simulations cannot be used to mimic observed data. We present novel rate equations model to capture the complex dynamics of these ego networks.

C 2: Adversarial/Trust

Joining User Profiles Across Online Social Networks: from the Perspective of an Adversary

Qiang Ma, Han Hee Song, S Muthukrishnan and Antonio Nucci

Being the anchor points for building social relationships in the cyber-space, online social networks (OSNs) play an integral part of modern peoples life. Since different OSNs are designed to address specific social needs, people take part in multiple OSNs to cover different facets of their life. While the fragmented pieces of information about a user in each OSN may be of limited use, serious privacy issues arise if a sophisticated adversary pieces information together from multiple OSNs. To this end, we undertake the role of such an adversary and demonstrate the possibility of "splicing" user profiles across multiple OSNs and present associated security risks to users. In doing so, we develop a scalable and systematic profile joining scheme, Splicer, that focuses on various aspects of profile attributes by simultaneously performing exact, quasiperfect and partial matches between pairs of profiles. From our evaluations on three real OSN data, Splicer not only handles large-scale OSN profiles efficiently by saving % computation time compared to all-pair profile comparisons, but also far exceeds the recall of generic distance measure based approach at the same precision level by %. Finally, we quantify the amount of information "lift" attributed to joining of OSNs, where on average % additional profile attributes can be added to % of users.

Prediction of Cyberbullying Incidents in a Media-based Social Network

Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv and Shivakant Mishra

Cyberbullying is a major problem affecting more than half of all American teens. Prior work has largely focused on detecting cyberbullying after the fact. In this paper, we investigate the prediction of cyberbullying incidents in Instagram, a popular media-based social network. The novelty of this work is building a predictor that can anticipate the occurrence of cyberbullying incidents before they happen. The Instagram media-based social network is well-suited to such prediction since there is an initial posting of an image typically with an associated text caption, followed later by the text comments that form the basis of a specific cyberbullying incident. We extract several important features from the initial posting data for automated cyberbullying prediction, including profanity and linguistic content of the text caption, image content, as well as social graph parameters and temporal content behavior. Evaluations using a real-world Instagram dataset demonstrate that our method achieves high performance in predicting the occurrence of cyberbullying incidents.

Hiding in Plain Sight: Characterizing and Detecting Malicious Facebook Pages

Prateek Dewan, Shrey Bagroy and Ponnurangam Kumaraguru

Facebook is the world's largest Online Social Network, having more than billion users. Like most other social networks, Facebook is home to various categories of hostile entities who abuse the platform by posting malicious content. In this paper, we identify and characterize Facebook pages that engage in spreading URLs pointing to malicious domains. We revisit the scope and definition of what is deemed as "malicious" in the modern day Internet, and identify pages publishing untrustworthy information, misleading content, adult and child unsafe content, scams, etc. Our findings revealed that at least % of all malicious pages were dedicated to promote a single malicious domain. Studying the temporal posting activity of pages revealed that malicious

pages were . times more active daily than benign pages. We further identified collusive behavior within a set of malicious pages spreading adult and pornographic content. Finally, we attempted to automate the process of detecting malicious Facebook pages by training multiple supervised learning algorithms on our dataset. Artificial neural networks trained on a fixed sized bag-of-words performed the best and achieved an accuracy of .%.

Detecting Malicious Campaigns in Crowdsourcing Platforms

Hongkyu Choi, Kyumin Lee and Steve Webb

Crowdsourcing systems enable new opportunities for requesters with limited funds to accomplish various tasks using human computation. However, the power of human computation is abused by malicious requesters who create malicious campaigns to manipulate information in web systems such as social networking sites, online review sites, and search engines. To mitigate the impact and reach of these malicious campaigns to targeted sites, we propose and evaluate a machine learning based classification approach for detecting malicious campaigns in crowdsourcing platforms as a first line of defense. Specifically, we (i) conduct a comprehensive analysis to understand the characteristics of malicious campaigns and legitimate campaigns in crowdsourcing platforms, (ii) propose various features to distinguish between malicious campaigns and legitimate campaigns, and (iii) evaluate a classification approach against baselines. Our experimental results show that our proposed approaches effectively detect malicious campaigns with low false negative and false positive rates.

Trust And Privacy Correlations in Social Networks: A Deep Learning Framework

Shatha Jaradat, Nima Dokoohaki, Mihhail Matskin and Elena Ferrari

Online Social Networks (OSNs) remain the focal point of Internet usage. Since the beginning, networking sites tried best to have right privacy mechanisms in place for users, enabling them to share the right content with the right audience. With all these efforts, privacy customizations remain hard for users across the sites. Existing research that address this problem mainly focus on semi-supervised strategies that introduce extra complexity by requiring the user to manually specify initial privacy preferences for their friends. In this work, we suggest an adaptive solution that can dynamically generate privacy labels for users in OSNs. To this end, we introduce a deep reinforcement learning framework that targets two key problems in OSNs like Facebook: the exposure of users' interactions through the network to less trusted direct friends, and the possibility of propagating user updates through direct friends' interactions to indirect friends. By implementing this framework, we aim at understanding how social trust and privacy could be correlated, specifically in a dynamic fashion. We report the ranked dependence between the generated privacy labels and the estimated user trust values, which indicate the ability of the framework to identify the highly trusted users and share with them higher percentages of data.

A 3: Algorithmic methods

NIMBLECORE: A Space-efficient External Memory Algorithm for Estimating Core Numbers

Priya Govindan, Sucheta Soundarajan, Tina Eliassi-Rad and Christos Faloutsos

We address the problem of estimating core numbers of nodes by reading edges of a large graph stored in external memory. The core number of a node is the highest k-core in which the node participates. Core numbers are useful in many graph mining tasks, especially ones that involve finding communities of nodes, influential spreaders and dense subgraphs. Large graphs often do not fit on the memory of a single machine. Existing external memory solutions do not give bounds on the required space. In practice, existing solutions

also do not scale with the size of the graph. We propose NimbleCore, an iterative externalmemory algorithm, which estimates core numbers of nodes using O(n log dmax) space, where n is the number of nodes and dmax is the maximum node-degree in the graph. We also show that NimbleCore requires O(n) space for graphs with power-law degree distributions. Experiments on forty-eight large graphs from various domains demonstrate that NimbleCore gives space savings up to X, while accurately estimating core numbers with average relative error less than .%.

All-Pairs Shortest Distances Maintenance in Relational DBMSs

Sergio Greco, Cristian Molinaro, Chiara Pulice and Ximena Quintana

Computing shortest distances is a central task in many graph applications. Although many algorithms to solve this problem have been proposed, they are designed to work in the main memory and/or with static graphs, which limits their applicability to many current applications where graphs are subject to frequent updates. In this paper, we propose novel efficient incremental algorithms for maintaining all-pairs shortest distances in dynamic graphs. We experimentally evaluate our approach on real-world datasets, showing that it outperforms current algorithms designed for the same problem.

Togetherness: An Algorithmic Approach to Network Integration

Anastasia Moskvina and Jiamou Liu

Network integration refers to a process of building links between two networks so that they dissolve into a single unified network. Togetherness measures the proximity of these two networks as they integrate; this notion is fundamental to social networks as it is relevant to important concepts such as trust, coherence and solidarity. In this paper, we study the algorithmic nature of network integration and formally introduce three notions of togetherness. We analyze the corresponding computational problems of network integration: Given two networks and a desired level of togetherness, build links between members of these networks so that the overall network meets the togetherness criterion. We analyze optimal solutions to this problem, describe several heuristics and compare their performance through experimental analysis.

On the Guarantee of Containment Probability in Influence Minimization

Chien-Wei Chang, Mi-Yen Yeh and Kun-Ta Chuang

We in this paper explore a novel model of influence minimization for the need to effectively prevent the outbreak of epidemicprone spread on networks. The current network-blocking models usually report the expected number of infected nodes under the limited number of cutting edges. However, to control the epidemic-prone spread such as dengue fever, epidemiologists tend to deploy a cost-effective intervention with low outbreak risk, but the outbreak risk cannot be estimated based on the expectation of infected count. We in this paper explore the first solution to estimate the probability that can successfully bound the infected count below the out-of-control threshold, which can be logically mapped to the outbreak risk and can facilitate the authority to adaptively adjust the intervention cost for the need of risk control. We elaborate upon the proposed MCP (standing for Maximization of Containment Probability) problem and show that it is a NP-hard challenge without the submodular property. We further devise an effective measurement of sufficient number of Monte Carlo iterations based on the relative error of Monte Carol integration. The experimental results show that our proposed algorithm with small iterations can deliver the qualified guarantee of containment probability, demonstrating its feasibility for real applications.

B 3: Information Sharing

Intertwined Viral Marketing in Social Networks

Jiawei Zhang, Senzhang Wang, Qianyi Zhan and Philip S Yu

Traditional viral marketing problems aim at selecting a subset of seed users for one single product to maximize its awareness in social networks. However, in real scenarios, multiple products can be promoted in social networks at the same time. At the product level, the relationships among these products can be quite intertwined, e.g., competing, complementary and independent. In this paper, we will study the "interTwined Influence Maximization" (i.e., TIM) problem for one product that we target on in online social networks, where multiple other competing/complementary/independent products are being promoted simultaneously. The TIM problem is very challenging to solve due to () few existing models can handle the intertwined diffusion procedure of multiple products concurrently, and () optimal seed user selection for the target product may depend on other products' marketing strategies a lot. To address the TIM problem, a unified greedy framework TIER (interTwined Influence EstimatoR) is proposed in this paper. Extensive experiments conducted on four different types of realworld social networks demonstrate that TIER can outperform all the comparison methods with significant advantages in solving the TIM problem.

Analyzing information sharing strategies of users in online social networks

Dong-Anh Nguyen, Shulong Tan, Ram Ramanathan and Xifeng Yan

User information sharing is an important behavior in online social networks. Understanding such behavior could help in various applications such as user modeling, information cascade analysis, viral marketing, etc. In this paper, we aim to understand the strategies users employ to make retweet decision. We are interested in investigating whether these strategies in online social network contain significant information about users and can be used to further characterize users. We propose a flexible model that captures a number of behavior signals affecting user's retweet decision. Our empirical results show that the inferred strategies can help increase the performance of retweet prediction.

Learning Cascaded Influence under Partial Monitoring

Jie Zhang, Jiaqi Ma and Jie Tang

Social influence has attracted tremendous attention from both academic and industrial communities due to the rapid development of online social networks. While most research has been focused on the direct influence between peers, learning cascaded indirect influence has not been previously studied. In this paper, we formulate the concept of cascade indirect influence based on the Independent Cascade model and then propose a novel online learning algorithm for learning the cascaded influence in the partial monitoring setting.We propose two bandit algorithms E-EXP and RE-EXP to address this problem. We theoretically prove that E-EXP has a cumulative regret bound of O(pT) over T, the number of time stamps. We will also show that RE-EXP, a relaxed version of E-EXP, achieves a better performance in practice. We compare the proposed algorithms with three baseline methods on both synthetic and real networks (Weibo and AMiner). Our experimental results show that RE-EXP converges faster than E-EXP. Both of them significantly outperform the alternative methods in terms of normalized regret. Finally, we apply the learned cascaded influence to help behavior prediction and experiments show that our proposed algorithms can help achieve a significant improvement (-% by accuracy) for behavior prediction.

Community-based delurking in social networks

Roberto Interdonato, Chiara Pulice and Andrea Tagarelli

The participation inequality phenomenon in online social networks between the niche of super contributors and the crowd of silent users, a.k.a. lurkers, has been witnessed in many domains. Within this view, understanding the role that lurkers take in the network is essential to develop innovative strategies to delurk them, i.e., to engage such users into a more active participation in the social network life. In this work, we leverage the boundary spanning theory to enhance our understanding of lurking behaviors, with the goal of improving the task of delurking in social networks. Assuming the availability of a global community structure, we first analyze how lurkers are related to users that take the role of bridges between different communities, unveiling insights into the bridging nature of lurkers and their tendency to acquire information from outside their own community. Moreover, based on a targeted influence maximization method designed for delurking, we also analyze how the learning of users that can best engage lurkers is related to the community structure. We found that the best users to engage lurkers belonging to any particular community, are more often found outside that community, and more specifically they are located in the adjacent communities.

C 3: Spatial

Percimo: A Personalized Community Model for Location Estimation in Social Media

Guangchao Yuan, Pradeep Kumar Murukannaiah and Munindar Singh

User location is crucial in understanding the dynamics of user activities, especially in relating their online and offline aspects. However, users' social media activities, such as tweets sent, do not always reveal their location. We consider the problem of estimating geo-tags for tweets and develop a comprehensive approach that incorporates textual content, the user's personalized behavior, and the user's social relationships. Our approach, Percimo, considers the two major kinds of communal attachment, which have distinct computational ramifications. We evaluate Percimo via three geo-social graphs based on the mutual-follow relationships of Twitter users, their geographical distance (computed from their geo-tagged tweets), and their preferences for location categories (collected from Foursquare). We find that Percimo yields a smaller prediction error than the two state-of-the-art approaches we compare with.

Community-Based Geospatial Tag Estimation

Wei Niu, James Caverlee, Haokai Lu and Krishna Kamath

This paper tackles the geospatial tag estimation problem, which is of critical importance for location-based search, retrieval, and mining applications. However, tag estimation is challenging due to massive sparsity, uncertainty in the tags actually used, as well as diversity across locations and times. Toward overcoming these challenges, we propose a communitybased smoothing approach that seeks to uncover hidden conceptual communities which link multiple related locations by their common interests in addition to their proximity. Through extensive experiments over a sample of millions of geo-tagged Twitter posts, we demonstrate the effectiveness of the smoothing approach and validate the intuition that geo-locations have the tendency to share similar "ideas" in the formation of conceptual communities.

Exploiting Spatial-Temporal-Social Constraints for Localness Inference Using Online Social Media

Chao Huang and Dong Wang

The localness inference problem is to identifyz whether a person is a local resident in a city or not and the likelihood of a venue to attract local people. This information is critical for many applications such as targeted ads of local business, urban planning, localized news and travel recommendations. While there are prior work

on geo-locating people in a city using supervised learning approaches, the accuracy of those techniques largely depends on a high quality training dataset, which is difficult and expensive to obtain in practice. In this study, we propose to exploit spatial-temporalsocial constraints from noisy online social media data to solve the localness inference problem using an unsupervised approach. The spatial-temporal constraint represents the correlations between people and venues they visit and the social constraint represents social connections between people. In particular, we develop a Spatial-Temporal-Social-Aware (STSA) inference framework to jointly infer i) the localness of a person and ii) the local attractiveness of a venue without requiring any training data. We evaluate the performance of STSA scheme using three real-world datasets collected from Foursquare. Experimental results show that STSA scheme outperforms the state-of-the-art techniques by significantly improving the estimation accuracy.

Co-Location Social Networks: Linking the Physical World and Cyberspace

Huandong Wang, Yong Li, Yang Chen, Yue Wang, Jian Yuan and Depeng Jin

Various dedicated web services in the cyberspace, e.g., social networks, e-commerce, and instant communications, play a significant role in people's daily-life. Billions of people around the world access them through multiple online identifiers (IDs), and interact with each other in both the cyberspace and the physical world. These two kinds of interactions are highly relevant to each other. In order to link between the cyberspace and the physical world, we propose a new type of social network, i.e., co-location social network (CLSN). A CLSN contains online IDs describing people's online presence and offline interactions when people come across each other. By analyzing real data collected from a mainstream ISP in China, which contains . million IDs across most popular web services, we build a largescale CLSN, and evaluate its unique properties. The results verify that the CLSN is quite different from existing online and offline social networks in terms of different classic graph metrics. This paper is the first research to study CLSN at scale and paves the way for future studies of this new type of social network.

A 4: Machine learning methods

Network Classification Using Adjacency Matrix Embeddings and Deep Learning

Ke Wu, Philip Watters and Malik Magdon-Ismail

We study a natural problem: Given a small piece of a large parent network, is it possible to identify the parent network? We approach this problem from two perspectives. First, using several "sophisticated" or "classical" network features that have been developed over decades of social network study. These features measure aggregate properties of the network and have been found to take on distinctive values for different types of network, at the large scale. By using these classical features within a standard machine learning framework, we show that one can identify large parent networks from small (even -node) subgraphs. Second, we present a novel adjacency matrix embedding technique which converts the small piece of the network into an image and, within a deep learning framework, we are able to obtain prediction accuracies upward of %, which is comparable to or slightly better than the performance from classical features. Our approach provides a new tool for topology-based prediction which may be of interest in other network settings. Our approach is plug and play, and can be used by non-domain experts. It is an appealing alternative to the often arduous task of creating domain specific features using domain expertise.

Weakly Hierarchical Lasso based Learning to Rank in Best Answer Prediction

Qiongjie Tian and Baoxin Li

In community question and answering sites, pairs of questions and their high-quality answers (like best answers selected by askers) can be valuable knowledge available to others. However lots of questions receive multiple answers but askers do not label either one as the accepted or best one even when some replies answer their questions. To solve this problem, highquality answer prediction or best answer prediction has been one of important topics in social media. These user-generated answers often consist of multiple "views", each capturing different (albeit related) information (e.g., expertise of the asker, length of the answer, etc.). Such views interact with each other in complex manners that should carry a lot of information for distinguishing a potential best answer from others. Little existing work has exploited such interactions for better prediction. To explicitly model these information, we propose a new learningto- rank method, ranking support vector machine (RankSVM) with weakly hierarchical lasso in this paper. The evaluation of the approach was done using data from Stack Overflow. Experimental results demonstrate that the proposed approach has superior performance compared with approaches in state-ofthe- art.

Priority Rank Model for Social Network Generation

Mikolaj Morzy, Przemyslaw Kazienko and Tomasz Kajdanowicz

Currently available artificial network generation models are characterized by consistency and low variance due to the rigidity of models' underlying assumptions. Networks generated from these models are usually too regular and do not contain noise and imbalance inherent in networks induced by human behavior. An important consequence is that much research on social network analysis presented in recent years used idealistic artificial networks that did not conform to reality. In order to alleviate this problem we introduce a new network generation model capable of modeling a broad spectrum of networks. In our model, the network formation process is not hard-coded into the model. Rather, we propose a simple mechanism for network creation based on priority ranking, and we encode the guiding principle of network formation as a distance function. By only changing the distance function definition and using the same priority ranking mechanism we are able to model very diverse networks. Our preliminary results show that we can mimic the behavior of popular artificial network generation models, such as the Erdös- Rényi random network model, the Watts-Strogatz small world model, or the Albert-Barabási preferential attachment model, but we can generate new types of networks as well. Following the principles of Open Science we publish the source code used to perform experiments and publish results in a public repository

An Information Theoretic Approach to Generalised Blockmodelling for the Identification of Meso-Scale Structure in Networks

Neil Hurley and Erika Duriakova

Blockmodelling is a technique whose aim is to identify meaningful structure in networks. Community finding is a type of blockmodelling in so far as it focuses on identifying dense subgraph structure. Generalised blockmodelling allows an analyst to explicitly control the type of extracted structure. When compared to the well studied community-finding problem, generalised blockmodelling algorithms lag well behind in terms of their scalability. In this paper we formulate and evaluate a generalised blockmodelling algorithm, based on the Infomap information-theoretic community-finding algorithm.We reformulate the optimisation objective of the Infomap algorithm, so that it is extended to identify specific types of meso-scale structure that are given as input by the analyst. We evaluate our method against other generalised blockmodelling algorithms.

Bayesian Model Selection of Stochastic Block Models

Xiaoran Yan

A central problem in analyzing networks is partitioning them into modules or communities. One of the best tools for this is the stochastic block model, which clusters vertices into blocks with statistically homogeneous

pattern of links. Despite its flexibility and popularity, there has been a lack of principled statistical model selection criteria for the stochastic block model. Here we propose a Bayesian framework for choosing the number of blocks as well as comparing it to the more elaborate degreecorrected block models, ultimately leading to a universal model selection framework capable of comparing multiple modeling combinations. We will also investigate its theoretic connection to the minimum description length principle.

B 4: Communities

Dynamic Community Detection based on Distance Dynamics

Qian Guo, Lei Zhang, Bin Wu and Xuelin Zeng

Dynamic community detection has been of great significance on analyzing network structure and community evolution. Among state-of-the-art methods, incremental algorithms based on modularity have been used widely, for the fully utilization of both current and historical information. Unfortunately, they are difficult to uncover small community due to problem called "resolution limit" and also sensitive to the sequence of network increments' arrival. In this paper, we propose a novel dynamic community detection algorithm based on distance dynamics, which detects community in near-linear time. Meanwhile, the proposed algorithm overcomes the shortcomings of traditional methods by replacing modularity with local interaction model. In a sense, it is assumed that increments can be treated as disturbance of network. It can be limited to a certain "local area" by disturbing factor. Experiments show that the proposed algorithm has achieved well balance between efficiency and effectiveness both in synthetic and real world networks.

Enumerating Maximal Cliques in Temporal Graphs

Anne-Sophie Himmel, Hendrik Molter, Rolf Niedermeier and Manuel Sorge

Dynamics of interactions play an increasingly important role in the analysis of complex networks. A modeling framework to capture this are temporal graphs. We focus on enumerating -cliques, an extension of the concept of cliques to temporal graphs: for a given time period , a -clique in a temporal graph is a set of vertices and a time interval such that all vertices interact with each other at least after every time steps within the time interval. Viard, Latapy, and Magnien [ASONAM] proposed a greedy algorithm for enumerating all maximal -cliques in temporal graphs. In contrast to this approach, we adapt to the temporal setting the Bron-Kerbosch algorithm—an efficient, recursive backtracking algorithm which enumerates all maximal cliques in static graphs. We obtain encouraging results both in theory (concerning worstcase time analysis based on the parameter “-slice degeneracy” of the underlying graph) as well as in practice with experiments on real-world data. The latter culminates in a significant improvement for most interesting -values concerning running time in comparison with the algorithm of Viard, Latapy, and Magnien (typically two orders of magnitude).

Structural Measures of Clustering Quality on Graph Samples

Jianpeng Zhang, Yulong Pei, George Fletcher and Mykola Pechenizkiy

Due to the growing presence of large-scale and streaming graphs such as social networks, graph sampling and clustering play an important role in many real-world applications. One key aspect of graph clustering is the evaluation of cluster quality. However, little attention has been paid to evaluation measures for clustering quality on samples of graphs. As first steps towards appropriate evaluation of clustering methods on sampled graphs, in this work we present two novel evaluation measures for graph clustering called -precision and -recall. These measures effectively reflect the match quality of the clusters in the sampled graph with respect to the ground-truth clusters in the original graph. We show in extensive experiments on various benchmarks that our proposed metrics are practical and effective for graph clustering evaluation.

A Local Measure of Community Change in Dynamic Graphs

Anita Zakrzewska, Eisha Nathan, James Fairbanks and David A. Bader

In this work we present a new local, vertex-level measure of community change. Our measure detects vertices that change community membership due to the actions (edges) of a vertex itself and not only due to global community shifts. The local nature of our measure is important for analyzing real graphs because communities may change to a large degree from one snapshot in time to the next. Using both real and synthetic graphs, we compare our measure to an alternative, global approach. Both approaches detect community switching vertices in a synthetic example with little overall community change. However, when communities do not evolve smoothly over time, the global approach flags a very large number of vertices, while our local method does not.

A Parameter-Free Method for Detecting Local Communities Based on Attainable Information

Ardavan Afshar, Mansour Zolghadri Jahromi and Ali Hamzeh

Community detection approaches are useful tools for revealing the structure and properties of networks. There are many approaches for identifying communities, some require inaccessible information, others need initializing parameters to perform well. Many existing algorithms require centralized decision maker to reveal communities. This paper proposes a non-centralized parameter-free method that only uses attainable information. The proposed method is examined using real as well as synthetic social networks. Experimental results suggest that this approach can perform as well as centralized approaches that require global information for detecting communities.

C 4: Applications

Web User Profiling using Data Redundancy

Xiaotao Gu, Hong Yang, Jie Tang and Jing Zhang

The study of Web user profiling can be traced back to years ago, with the goal of extracting “semantic”-based user profile attributes from the unstructured Web. Despite slight differences, the general method is to first identify relevant pages of a specific user and then use machine learning models (e.g., CRFs) to extract the profile attributes from the page. However, with the rapid growth of the Web volume, such a method suffers from data redundancy and error propagation between the two steps. In this paper, we revisit the problem of Web user profiling in the big data era, trying to deal with the new challenges. We propose a simple but very effective approach for extracting user profile attributes from the Web using big data. To avoid error propagation, the approach processes all the extraction subtasks in one unified model. To further incorporate human knowledge to improve the extraction performance, we propose a markov logic factor graph (MagicFG) model. The MagicFG model describes human knowledge as first-order logics and combines the logics into the extraction model. Our experiments on a real data set show that the proposed method significantly improves (+- %; p ≪ :, t-test) the extraction performance in comparison with several baseline methods.

Can I Foresee the Success of My Meetup Group?

Soumajit Pramanik, Midhun Gundapuneni, Sayan Pathak and Bivas Mitra

Success of Meetup groups is of utmost importance for the members who organize them. Given a wide variety of such groups, a single metric may not be indicative of success for different groups; rather, success measure should be specific to the interest of a group. In this paper, accounting for the group diversity, we

systematically define Meetup group success metrics and use them to generate labels for our machine learnt models. We crawl the Meetup dataset for three US cities namely New York, Chicago and San Francisco over a period of months. The data study reveals the key players (such as core members, new members etc.) behind the success of the Meetup groups. This study leverages semantic, syntactic, temporal and location based features to discriminate between successful and unsuccessful groups. Finally, we present a model to predict success of the Meetup groups with high accuracy (. with AUC = .). Our approach generalizes well across groups, categories and cities. Additionally, the model performs reasonably well for new groups with little history (cold start problem), exhibiting high accuracy for the cross city validation.

Subconscious Crowdsourcing: A Feasible Data Collection Mechanism for Mental Disorder Detection on Social Media

Chun-Hao Chang, Elvis Saravia and Yi-Shin Chen

Mental disorders are currently affecting millions of people from different cultures, age groups and geographic regions. The challenge of mental disorders is that they are difficult to detect on suffering patients, thus presenting an alarming number of undetected cases and misdiagnosis. In this paper, we aim at building predictive models that leverage language and behavioral patterns, used particularly in social media, to determine whether a user is suffering from two cases of mental disorder. These predictive models are made possible by employing a novel data collection process, coined as Subconscious Crowdsourcing, which helps to collect a faster and more reliable dataset of patients. Our experiments suggest that extracting specific language patterns and social interaction features from reliable patient datasets can greatly contribute to further analysis and detection of mental disorders.

An Analysis of Student Behavior in Two Massive Open Online Courses

James Schaffer, Brandon Huynh, John O'Donovan, Tobias Hollerer, Yinglong Xia and Sabrina Lin

Personality Homophily and the Local Network Characteristics of Facebook

Nyala Noe, Roger M. Whitaker and Stuart M. Allen

Social networks are known to form on the basis of homophily, where nodes with some type of similar characteristics are more likely to be connected. Some of the most fundamental human characteristics are reflected by an individual’s personality, which represents a persistent disposition governing a human’s outlook and approach to diverse situations. While taking into account demographics of age and gender, we assess the extent to which personality homophily is evident in the local network features of Facebook. Using a large sample obtained from the MyPersonality dataset, we find that a range of network-based features correlate with personality facets of individuals. In particular, extraversion had a positive effect on an individual’s network size, while neuroticism had a negative effect. Additionally, extraversion and openness were positively related to transitivity, which was moderated by gender. Finally, we found that conscientiousness, agreeableness and extraversion were homophilous: people with higher similarity on these facets were more strongly connected. This was additionally mediated by gender for agreeableness: personality similarity had an effect for male-only and mixed pairs, but not for female-only pairs. Personality similarity was also stronger among closed triangles, compared to open ones. These results support the idea that inherent attraction between individuals, on the basis of personality, drives the roles we play within our online social networks.

A 5: Recommender sys.

Downside Management in Recommender Systems

Huan Gui, Haishan Liu, Xiangrui Meng, Anmol Bhasin and Jiawei Han

In recommender systems, bad recommendations can lead to a net utility loss for both users and content providers. The downside (individual loss) management is a crucial and important problem, but has long been ignored. We propose a method to identify bad recommendations by modeling the users' latent preferences that are yet to be captured using a residual model, which can be applied independently on top of existing recommendation algorithms. We include two components in the residual utility: benefit and cost, which can be learned simultaneously from users' observed interactions with the recommender system. We further classify user behavior into fine-grained categories, based on which an efficient optimization algorithm to estimate the benefit and cost using Bayesian partial order is proposed. By accurately calculating the utility users obtained from recommendations based on the benefit-cost analysis, we can infer the optimal threshold to determine the downside portion of the recommender system. We validate the proposed method by experimenting with real-world datasets and demonstrate that it can help to prevent bad recommendations from showing.

Collaborative Restricted Boltzmann Machine for Social Event Recommendation

Xiaowei Jia, Xiaoyi Li, Kang Li, Vishrawas Gopalakrishnan, Guangxu Xun and Aidong Zhang

The development of social networks has not only improved the online experience, but also stimulated the advances in knowledge mining so as to assist people in planning their offline social events. Users can explore their favorite events, such as celebrations and symposiums, through the pictures and the posts from their friends on social networks. An effective event recommendation can offer great convenience for both event organizers and participants, which yet remains extremely challenging due to a wide range of practical concerns. In this paper we propose a novel recommendation framework, which combines the information from multiple sources and establishes a connection between the online knowledge and the event participation.

Twitter Message Recommendation Based on User Interest Profiles

Raheleh Makki, Axel J. Soto, Stephen Brooks and Evangelos E. Milios

Twitter has become one of the most important platforms for gathering information, where users follow breaking news, track ongoing events and learn about their topics of interest. Considering the sheer volume of Twitter data and the ever-growing number of users, it is of great importance to have real-time systems that can monitor and recommend relevant and non-redundant tweets with respect to users' interests. In this paper, we propose a framework using language models as a basis for analyzing strategies and techniques for tweet recommendation based on user interest profiles. Results show that identifying named entities in profiles has a major impact on the accuracy of the recommender. We also performed a thorough comparison to investigate whether state-of-the-art semantic relatedness techniques have a positive impact on the precision of the recommended tweets. The TREC Microblog track dataset is used for comparison and evaluation throughout this paper.

Generating Risk Reduction Recommendations to Decrease Vulnerability of Public Online Profiles

Janet Zhu, Sicong Zhang, Lisa Singh, Grace Hui Yang and Micah Sherr

Preserving online privacy is becoming increasingly challenging due in large part to the continued growth of social media. Those who choose to share their information publicly may not realize what features of their profiles make their public data more identifiable and potentially vulnerable to cross-site record linkage. This paper proposes a risk reduction recommendation method that suggests removal or modification of a small number of attributes to make a profile less unique, thereby reducing the identifiability and vulnerability of the

user. Empirical results on data collected from Google+, LinkedIn, and Foursquare show that users' vulnerability in terms of identifiability and data exposure level can be significantly reduced while public profile utility can be maintained using our proposed approach.

Exploring Influence Among Participants for Event Recommendation

Yi Liao, Xinshi Lin and Wai Lam

Event-based Social Networks (EBSN) are popular for organizing offline social events nowadays. In this paper, we develop a new model for event recommendation on EBSNs, which exploits the influence of existing participants, who have expressed willingness to join, on new participants in addition to other context information. Utilizing the participant influence can improve the effectiveness of event recommendation. Experiments on real datasets confirm that the consideration of participant influence can lead to more accurate prediction, offering better event recommendation.

B 5: Social Media

From Migration Corridors to Clusters: The Value of Google+ Data for Migration Studies

Johnnatan Messias, Fabricio Benevenuto, Ingmar Weber and Emilio Zagheni

Recently, there have been considerable efforts to use online data to investigate international migration. These efforts show that Web data are valuable for estimating migration rates and are relatively easy to obtain. However, existing studies have only investigated flows of people along migration corridors, i.e. between pairs of countries. In our work, we use data about "places lived" from millions of Google+ users in order to study migration 'clusters', i.e. groups of countries in which individuals have lived sequentially. For the first time, we consider information about more than two countries people have lived in. We argue that these data are very valuable because this type of information is not available in traditional demographic sources which record country-to-country migration flows independent of each other. We show that migration clusters of country triads cannot be identified using information about bilateral flows alone. To demonstrate the additional insights that can be gained by using data about migration clusters, we first develop a model that tries to predict the prevalence of a given triad using only data about its constituent pairs. We then inspect the groups of three countries which are more or less prominent, compared to what we would expect based on bilateral flows alone. Next, we identify a set of features such as a shared language or colonial ties that explain which triple of country pairs are more or less likely to be clustered when looking at country triples. Then we select and contrast a few cases of clusters that provide some qualitative information about what our data set shows. The type of data that we use is potentially available for a number of social media services. We hope that this first study about migration clusters will stimulate the use of Web data for the development of new theories of international migration that could not be tested appropriately before.

How Fashionable is Each Street?: Quantifying Road Characteristics using Social Media

Takuya Nishimura, Kyosuke Nishida, Hiroyuki Toda and Hiroshi Sawada

Determining routes that provide opportunities to satisfy the various demands of users is still an open problem. This is because it is virtually impossible to manually quantify the characteristics of each road and there are few resources describing roads directly such that we meet any demand that may arise. The goal of this study is to automatically quantify the characteristics of roads for demands that can be described using keywords such as "fashionable". To achieve this goal, we propose a two-stage method that analyzes social media and road networks. First, our method estimates the topic distribution (i.e., the characteristics) of each point-of-interest (POI) by analyzing geo-tagged texts with the Latent Dirichlet Allocation model. Next, it uses a Markov

random field model to estimate the characteristics of each road on the basis of those of POIs and the road networks associated with the POIs. Experiments on real datasets demonstrate that our method achieves statistically significant improvements over baseline methods in terms of ranking quality in the information retrieval for roads in three areas given keywords.

From Event Detection to Storytelling on Microblogs

Janani Kalyanam, Sumithra Velupillai, Mike Conway and Gert Lanckriet

The problem of detecting events from content published on microblogs has garnered much interest in recent times. In this paper, we address the questions of what happens after the outbreak of an event in terms of how the event gradually progresses and attains each of its milestones, and how it eventually dissipates. We propose a model based approach to capture the gradual unfolding of an event over time. This enables the model to automatically produce entire timeline trajectories of events from the time of their outbreak to their disappearance. We apply our model on the Twitter messages collected about Ebola during the outbreak and obtain the progression timelines of several events that occurred during the outbreak. We also compare our model to several existing topic modeling and event detection baselines in literature to demonstrate its efficiency.

Authorship Identification in Bengali Language: A Graph Based Approach

Tanmoy Chakraborty and Prasenjit Choudhury

Individuals have distinctive ways of speaking and writing, and there exists a long history of linguistic and stylistic investigation into authorship attribution. Most authorship identification approaches are exclusively based on lexical measures such as vocabulary richness and lexico-syntactic features, or substantially generate relevant features for different machine learning approaches. These techniques are inefficient without suitable feature selection or large corpus. In this paper, we introduce three graph based models for the task of authorship identification, that consider the interaction of character sequences, phraseological patterns and structure of the sentences in the document to construct graphs separately for an author. For each model, all such graphs for different authors are aggregated to generate a combined weighted training graph. Then a simple graph traversal algorithm is used to compare testing graph with the training graph. The experiment is conducted on the documents of six authors collected from Bengali literature. Experimental results show that our models significantly outperform four state-of-theart models (.% higher than the best baseline model) even for short training dataset.

Finding Needles of Interested Tweets in the Haystack of Twitter Network

Qiongjie Tian, Jashmi Lagisetty and Baoxin Li

Drug use and abuse is a serious societal problem. The fast development and adoption of social media and smart mobile devices in recent years bring about new opportunities for advancing computer-based strategies for understanding and intervention of drug-related behaviors. However, the existing literature still lacks principled ways of building computational models for supporting effective analysis of large-scale, often unstructured social media data. Part of the challenge stems from the difficulty of obtaining so-called ground-truth data that are typically required for training computational models. This paper presents a progressive semi-supervised learning approach to identifying Twitter tweets that are related to personal and recreational use of marijuana. Based on a small, labeled dataset, the proposed approach first learns optimal mapping of raw features from the tweets for classification, using a method of weakly hierarchical lasso. The learned feature model is then used to support unsupervised clustering of Web-scale data. Experiments with realistic data crawled from Twitter are used to validate the proposed approach, demonstrating its effectiveness.

C 5: New Directions

Social Badge System Analysis

Jiawei Zhang, Xiangnan Kong and Philip S. Yu

To incentivize users' participations, online social networks often provide users with various rewards for their contributions to the sites. Attracted by the rewards, users will spend more time using the network services. Specifically, in this paper, we will mainly focus on "badges reward systems". Badges are small icons attached to users' homepages and profiles denoting their achievements. People like to accumulate badge for various reasons, which are modeled as the "badge values" in this paper. Meanwhile, to get badges, people also need to exert efforts to finish the required tasks, which will lead to certain "costs" as well. To understand users' badge achievement activities better, we will study an existing badge system launched in a real-world online social network, Foursquare, in this paper. A longer version of this paper is available at [].

A First Look at User Activity on Tinder

Gareth Tyson, Vasile Claudiu Perta, Hamed Haddadi and Micheal Seto

Mobile dating apps have become a popular means to meet potential partners. Although several exist, one recent addition stands out amongst all others. Tinder presents its users with pictures of people geographically nearby, whom they can either like or dislike based on first impressions. If two users like each other, they are allowed to initiate a conversation via the chat feature. In this paper we use a set of curated profiles to explore the behaviour of men and women in Tinder. We reveal differences between the way men and women interact with the app, highlighting the strategies employed. Women attain large numbers of matches rapidly, whilst men only slowly accumulate matches. Most notably, our results indicate that a little effort in grooming profiles, especially for male users, goes a long way in attracting attention.

New to Online Dating? Learning from Experienced Users for a Successful Match

Mo Yu, Xiaolong Zhang and Derek Kreager

Online dating arises as a popular venue for finding romantic partners in recent years. Many online dating sites adopt recommender systems to help their users. However, few of current research provides solutions to cold start problem, i.e., providing recommendations to new users. In this research, we propose a new approach of providing reciprocal online dating recommendations to new users. Specifically, we detect communities from existing users, match new users to these communities, and take advantage of reciprocal activities of those community members to provide recommendations to new users. Using data from a popular U.S. online dating site, experiments show that our approach greatly outperforms existing methods.

Dynamics of large scale networks following a merger

John Clements, Babak Farzad and Henryk Fuks

We studied the dynamic network of relationships among avatars in the massively multiplayer online game Planetside . In the spring of , two separate servers of this game were merged, and as a result, two previously distinct networks were combined into one. We observed the evolution of this network in the seven month period following the merger. We found that some structures of original networks persist in the combined network for a long time after the merger. As the original avatars are gradually removed, these structures slowly dissolve, but they remain observable for a surprisingly long time. We present a number of visualizations illustrating the postmerger dynamics and discuss time evolution of selected quantities characterizing the topology of the network.

The Collapse of the Friendster Network Started From the Center of the Core

Kazunori Seki and Masataka Nakamura

Friendster is a social networking service which used to be popular all over the world at the beginning of the st century and declined thereafter. Some researchers have examined the process of decline and explained that the social network on Friendster collapsed from the outside of the core structure. In order to verify if their assertion is true, we analyze the time evolution of the network structure more carefully. The result implies that the collapse of the Friendster networks actually started from the center of the core structure. We also attempt to explain its mechanism by a model based on the SIR model, which is a model in epidemiology. The findings imply that the tendency of "non-users who have many friends on Friendster are likely to register for Friendster" and "users who have many friends that are already tired of Friendster are likely to leave Friendster" played an important role in the time evolution of the core structure of the network. The tendency of "users who have few friends on Friendster are likely to leave soon" would have also played a key role in the process.

A 6: Markets

Targeting Algorithms for Online Social Advertising Markets

Chaolun Xia, Saikat Guha and Shan Muthukrishnan

Advertisers in online social networks (OSNs) like Facebook and LinkedIn have some preferred set of users they wish to reach by showing their ads. OSNs offer fine-grained sets of user characteristics – including their career, wealth, education information, etc — that advertisers can specify for targeting their audience, and each of these characteristics requires different amounts of money for targeting. The problem we address is what we call the targeting problem, that is, given a set ST characteristics of interest to an advertiser (that is, the advertiser wishes to reach users who have these characteristics, i.e. U(ST)) and a budget b he wishes to spend, how to split the budget among the user characteristics so that they can reach the most number of users in U(ST)? OSN-perspective. OSNs have complete knowledge of each user and their characteristics. In this case, we propose a polynomial time algorithm for the targeting problem and prove that it is an □=e approximation to the targeting that gets the optimal number of users. We define the marginal increment and iteratively maximize it. Advertiser-perspective. No single advertiser has the mapping of users to their characteristics or the distribution of users over the characteristics, and hence they cannot use the algorithm from above. We show through empirical data analysis that the strategy of targeting subsets U(S) U(ST) is their only feasible approach (in other words, targeting U(S) U(ST) will be arbitrarily worse than the direct solution). For evaluation, we crawl and analyze more than one million suggested bids from Facebook and LinkedIn. Further, we propose a fast greedy algorithm for the targeting problem based on targeting subsets, that in empirical analysis, increases the number of reached preferred users by nearly % over directly targeting the characteristics of interest, and for a moderate budget, it increases the number of reached preferred users by nearly %.

Correlations of consumption patterns in social-economic networks

Yannick Leo, Marton Karsai, Carlos Sarraute and Eric Fleury

We analyze a coupled dataset collecting the mobile phone communication and bank transactions history of a large number of individuals living in Mexico. After mapping the social structure and introducing indicators of socioeconomic status, demographic features, and purchasing habits of individuals we show that typical consumption patterns are strongly correlated with identified socioeconomic classes leading to patterns of

stratification in the social structure. In addition we measure correlations between merchant categories and introduce a correlation network, which emerges with a meaningful community structure. We detect multivariate relations between merchant categories and show correlations in purchasing habits of individuals. Our work provides novel and detailed insight into the relations between social and consuming behaviour with potential applications in recommendation system design.

Analysis of Rewards on Reward-Based Crowdfunding Platforms

Yusan Lin, Wang-Chien Lee and Chung-Chou H. Chang

Today, crowdfunding has emerged as a popular means for fundraising. Among various crowdfunding platforms, reward-based ones are the most well received. However, to the best knowledge of the authors, little research has been performed on rewards. In this paper, we analyze a Kickstarter dataset, which consists of approximately K projects and K rewards. The analysis employs various statistical methods, including Pearson correlation tests, Kolmogorov-Smirnow test and Kaplan-Meier estimation, to study the relationships between various reward characteristics and project success. We find that projects with more rewards, with limited offerings and late-added rewards are more likely to succeed.

Heuristics for Advertising Revenue Optimization in Online Social Networks

Inzamam Rahaman and Patrick Hosein

Recent increases in the adoption of Online Social Networks (OSNs) for advertising has resulted in the research and development of algorithms that can maximize the resulting revenue. OSN users are likely to be influenced by their friends; therefore, one can leverage friendship relationships to determine how advertisements should be distributed among users. If a user is given an indication that their friend clicked on an advertisement link (called an impression), then they are more likely to also click on the impression if it were to be provided to them. The problem of assigning impressions can be modeled as an optimization problem in which the goal is to maximize the expected number of clicks achieved given a fixed number of impressions. Hosein and Lawrence [] formulated this as a Stochastic Dynamic Programming problem in which impressions are provided in stages and the outcomes of previous stages are used in making impression allocations for the present stage. However, the determination of the optimal solution is computationally intractable for large problems; hence we require heuristics that are efficient while providing near-optimal solutions. In this paper we provide and compare various heuristics for this problem.

B 6: Social media

Emotion- and Area- Driven Topic Shift Analysis in Social Media Discussions

Kamil Topal, Mehmet Koyuturk and Gultekin Ozsoyoglu

Internet-based social media platforms allow individuals to discuss/comment on the ”topic” of an article in an interactive manner. The topic of a comment/reply in these discussions occasionally shifts, sometimes drastically and abruptly, other times slightly, away from the topic of the article. In this paper we study the phenomena of topic shifts in articleoriginated social media comments, and identify quantitatively the effects on topic shifts of comments (i) emotion levels (of various emotion dimensions), (ii) topic areas, and (iii) the structure of the discussion tree. We show that, with a better understanding of the topic shift phenomena in comments, automated systems can easily be built to personalize and cater to the comment-browsing and comment-viewing needs of different users.

What we write about when we write about causality: Features of causal statements across large-scale social discourse

Thomas McAndrew, Joshua Bongard, Chris Danforth, Peter Dodds, Paul Hines and James Bagrow

Identifying and communicating relationships between causes and effects is important for understanding our world, but is affected by language structure, cognitive and emotional biases, and the properties of the communication medium. Despite the increasing importance of social media, much remains unknown about causal statements made online. To study real-world causal attribution, we extract a large-scale corpus of causal statements made on the Twitter social network platform as well as a comparable random control corpus. We compare causal and control statements using statistical language and sentiment analysis tools. We find that causal statements have a number of significant lexical and grammatical differences compared with controls and tend to be more negative in sentiment than controls. Causal statements made online tend to focus on news and current events, medicine and health, or interpersonal relationships, as shown by topic models. By quantifying the features and potential biases of causality communication, this study improves our understanding of the accuracy of information and opinions found online.

Core-Periphery Clustering and Collaboration Networks

Pierluigi Crescenzi, Pierre Fraigniaud, Zvi Lotker and Paolo Penna

In this paper we analyse the core-periphery clustering properties of collaboration networks, where the core of a network is formed by the nodes with highest degree. In particular, we first observe that, even for random graph models aiming at matching the degree-distribution and/or the clustering coefficient of real networks, these models produce synthetic graphs which have a spatial distribution of the triangles with respect to the core and to the periphery which does not match the spatial distribution of the triangles in the real networks. We therefore propose a new model, called CPCL, whose aim is to distribute the triangles in a way fitting with their real coreperiphery distribution, and thus producing graphs matching the core-periphery clustering of real networks.

Temporal Mechanisms of Polarization in Online Reviews

Antonis Matakos and Panayiotis Tsaparas

In this paper we study the temporal evolution of review ratings. We observe that on average ratings tend to become more polarized over time. To explain this phenomenon we propose a simple model that captures the tendency of users for rating manipulation. Simulations with our model demonstrate that it is successful in capturing the aggregate behavior of the users.

C 6: Adversarial/Trust

A New Approach to Bot Detection: Striking the Balance Between Precision and Recall

Fred Morstatter, Liang Wu, Tahora Hossein Nazer, Kathleen Carley and Huan Liu

The presence of bots has been felt in many aspects of social media. Twitter, one example of social media, has especially felt the impact, with bots accounting for a large portion of its users. These bots have been used for malicious tasks such as spreading false information about political candidates and inflating the perceived popularity of celebrities. Furthermore, these bots can change the results of common analyses performed on social media. It is important that researchers and practitioners have tools in their arsenal to remove them. Approaches exist to remove bots, however they focus on precision to evaluate their model at the cost of recall. This means that while these approaches are almost always correct in the bots they delete, they ultimately delete very few, thus many bots remain. We propose a model which increases the recall in detecting bots,

allowing a researcher to delete more bots. We evaluate our model on two real-world social media datasets and show that our detection algorithm removes more bots from a dataset than current approaches.

Detecting Misinformation In Online Social Networks Before It Is Too Late

Huiling Zhang, Alan Kuhnle, Huiyuan Zhang and My T. Thai

Trustingness & Trustworthiness: A Pair of Complementary Trust Measures in a Social Network

Atanu Roy, Chandrima Sarkar, Jaideep Srivastava and Jisu Huh

The increase in analysis of real life social networks has led to a better understanding of the ways humans socialize in a group. Since trust is an important part of any social interaction, researchers use such networks to understand the nuances of trust relationships. One of the major requirements in trust applications is identifying the trustworthy actors in these networks. This paper proposes a pair of complementary measures that can be used to measure trust scores of actors in a social network using involvement of social networks. Based on the proposed measures, an iterative matrix convergence algorithm is developed that calculates the trustingness and the trustworthiness of each actor in the network. Trustingness of an actor is defined as the propensity of an actor to trust his neighbors in the network. Trustworthiness, on the other hand, is defined as the willingness of the network to trust an individual actor. The algorithm is proposed based on the idea that a person having higher trustingness score contributes to the trustworthiness of its neighbors to a lower degree. Conversely, a higher trustworthiness score is a result of lots of neighbors linked to the actor having low trustingness scores. The algorithm runs in O(k jEj) time where k denotes the number of iterations and jEj denotes the number of edges in the network. Moreover, the paper shows that the algorithm converges to a finite value quickly. Finally the proposed scores for trust prediction is implemented for various social networks and is shown that the proposed algorithm performs better (average %) than the state of the art trust scoring algorithms. The full version of the paper along with the coding implementation can be found at [].

ClearView: Data Cleaning for Online Review Mining

Amanda Minnich, Noor Abu-El-Rub, Maya Gokhale, Ronald Minnich and Abdullah Mueen

How can we automatically clean and curate online reviews to better mine them for knowledge discovery? Typical online reviews are full of noise and abnormalities, hindering semantic analysis and leading to a poor customer experience. Abnormalities include non-standard characters, unstructured punctuation, different/multiple languages, and misspelled words. Worse still, people will leave “junk” text, which is either completely nonsensical, spam, or fraudulent. In this paper, we describe three types of noisy and abnormal reviews, discuss methods to detect and filter them, and, finally, show the effectiveness of our cleaning process by improving the overall distributional characteristics of review datasets.

ASONAM Industrial Track Session 1: Knowledge from social and mobile media

Forecasting Price Shocks with Social Attention and Sentiment Analysis

Li Zhang, Liang Zhang, Keli Xiao and Qi Liu

Many recent studies on finance and social networks discovered that investor's attention is correlated to the financial market movement in terms of the price shocks. Following related findings, a significant and challenging problem is to forecast the direction of the market movement based on vast social media activities. Appropriately processing social networks data and developing models to capture investor's attention on stocks would effectively help financial forecasting. In this paper, we propose and then apply a price shocks forecasting framework, which simultaneously takes the influence of social network users and their opinions about stocks into consideration. Specifically, we develop a new method to estimate social attention to stocks by influence modeling and sentiment analysis. Then, we use it in price shocks forecasting, which we formalize as a classification problem. We also consider the effect of historical market information on the market movement. Finally, we evaluate our framework based on a series of tests on the Chinese stock data. Our results show that the newly proposed measurement of social attention effectively improves the forecasting power of our framework.

What's in the Community Cookie Jar?

Aaron Cahn, Scott Alfeld, Paul Barford and S Muthukrishnan

Third party tracking of user behavior via web cookies represents a privacy threat. In this paper we assess this threat through an analysis of anonymized, crowd-sourced cookie data provided by Cookiepedia.co.uk. We find that nearly % of the cookies in the corpus are from Facebook and of the remaining cookies % come from distinct domains. Over % are Maximal Permission cookies (i.e., rd party, non-secure, persistent, root-level). Cookiepedia's anonymization of user data presents challenges with respect to modeling site traffic. We further elucidate the privacy issue by conducting targeted crawling campaigns to supplement the Cookiepedia data. We find that the amount of traffic obscured by Cookiepedia's anonymizing procedure varies dramatically from site to site – sometimes obscuring as much as % of traffic. We use the crawls to infer the inverse function of the anonymizing procedure, allowing us to enhance the crowd-sourced dataset while maintaining user anonymity.

Mining Hidden Constrained Streams in Practice: Informed Search in Dynamic Filter Spaces

Nikolaos Panagiotou, Ioannis Katakis, Dimitrios Gunopulos, Vana Kalogeraki, Elizabeth Daly, Jia Yuan Yu and Brendan O!& Brien

In this paper we tackle the recently proposed problem of hidden streams. In many situations, the data stream that we are interested in, is not directly accessible. Instead, part of the data can be accessed only through applying filters (e.g. keyword filtering). In fact this is the case of the most discussed social stream today, Twitter. The problem in this case is how to retrieve as many relevant documents as possible by applying the most appropriate set of filters to the original stream and, at the same time, respect a number of constrains (e.g. maximum number of filters that can be applied). In this work we introduce a search approach on a dynamic filter space. We utilize heterogeneous filters (not only keywords) making no assumptions about the attributes of the individual filters. We advance current research by considering realistically hard constraints based on real-world scenarios that require tracking of multiple dynamic topics. We demonstrate the effectiveness of our approaches on a set of topics of static and dynamic nature. The development of the approach was motivated by a real application. Our system is deployed in Dublin City's Traffic Management Center and allows the city officers to analyze large sources of heterogeneous data and identify events related to traffic as well as emergencies.

A Bayesian Approach to Income Inference in a Communication Network

Martin Fixman, Ariel Berenstein, Jorge Brea, Martin Minnoni, Matias Travizano and Carlos Sarraute

The explosion of mobile phone communications in the last years occurs at a moment where data processing power increases exponentially. Thanks to those two changes in a global scale, the road has been opened to use mobile phone communications to generate inferences and characterizations of mobile phone users. In this work, we use the communications network, enriched by a set of users' attributes, to gain a better understanding of the demographic features of a population. Namely, we use call detail records and banking information to infer the income of each person in the graph.

ASONAM Industrial Track Session 2: Network, community, and cascades

Bridge the Terminology Gap Between Recruiters and Candidates: A Multilingual Skills Base built from Social Media and Linked Data

Emmanuel Malherbe and Marie-Aude Aufaure

A major part of the job offers and candidates profiles are now available online. Leveraging this public data, Multiposting, a subsidiary of SAP, aims at providing in realtime an exhaustive job market analysis through the SmartSearch project. One big issue in this project, and more generally in the e-recruitment and the human resources management, is to extract the skills from the raw texts in order to associate a job or a candidate to its corresponding skills. This paper proposes to generate a multilingual base of skills in a novel bottomup approach that finds its roots from the terminology used by candidates in professional social networks. The knowledge base is built by leveraging the Linked Open Data project DBpedia, as well as the tags of a Q&A website, StackOverflow. The large-scale experiments on real-world job offers show that the coverage and precision of the skills extraction are higher using this base than existing bases. The system has been implemented in industrial context and is used daily to extract the skills from thousands of documents, leading to advanced statistics as illustrated at the end this paper.

A Comparison of Methods for Cascade Prediction

Ruocheng Guo and Paulo Shakarian

Information cascades exist in a wide variety of platforms on Internet. A very important real-world problem is to identify which information cascades can "go viral". A system addressing this problem can be used in a variety of applications including public health, marketing and counter-terrorism. As a cascade can be considered as compound of the social network and the time series. However, in related literature where methods for solving the cascade prediction problem were proposed, the experimental settings were often limited to only a single metric for a specific problem formulation. Moreover, little attention was paid to the run time of those methods. In this paper, we first formulate the cascade prediction problem as both classification and regression. Then we compare three categories of cascade prediction methods: centrality based, feature based and point process based. We carry out the comparison through evaluation of the methods by both accuracy metrics and run time. The results show that feature based methods can outperform others in terms of prediction accuracy but suffer from heavy overhead especially for large datasets. While point process based methods can also run into issue of long run time when the model can not well adapt to the data. This paper seeks to address issues in order to allow developers of systems for social network analysis to select the most appropriate method for predicting viral information cascades.

AppVec: Vector Modeling of Mobile Apps and Applications

Qiang Ma, S. Muthukrishnan and Wil Simpson

We design a way to model apps as vectors, inspired by the recent deep learning approach to vectorization of words called wordvec. Our method relies on how users use apps. In particular, we visualize the time series of

how each user uses mobile apps as a “document”, and apply the recent wordvec modeling on these documents, but the novelty is that the training context is carefully weighted by the time interval between the usage of successive apps. This gives us the appvec vectorization of apps. We apply this to industrial scale data from Yahoo! and (a) show examples that appvec captures semantic relationships between apps, much as wordvec does with words, (b) show using Yahoo!’s extensive human evaluation system that % of the retrieved top similar apps are semantically relevant, achieving % lift over bag-of-word approach and % lift over matrix factorization approach to vectorizing apps, and (c) finally, we use appvec to predict app-install conversion and improve ad conversion prediction accuracy by almost %. This is the first industry scale design, training and use of app vectorization.

Analyzing the Spread of Chagas Disease with Mobile Phone Data

Juan de Monasterio, Alejo Salles, Carolina Lang, Diego Weinberg, Martin Minnoni, Matias Travizano and Carlos Sarraute

We use mobile phone records for the analysis of mobility patterns and the detection of possible risk zones of Chagas disease in two Latin American countries. We show that geolocalized call records are rich in social and individual information, which can be used to infer whether an individual has lived in an endemic area. We present two case studies, in Argentina and in Mexico, using data provided by mobile phone companies from each country. The risk maps that we generate can be used by health campaign managers to target specific areas and allocate resources more effectively.

ASONAM - Multidisciplinary Track S1: Applications of network analysis and social media analysis

Social Event Network Analysis: Structure, Preferences, and Reality

Martin Atzmueller, Tom Hanika, Gerd Stumme, Richard Schaller and Bernd Ludwig

This paper focuses on the analysis of socio-spatial data, i. e., user–performance relations at a distributed event. We consider the data as a bimodal network (i. e., model it as a bipartite graph), and investigate its structural characteristics towards a social network. We focus on plans of the participants (expressed by preferences) and their fulfilment, and propose measures for matching preference and reality. We specifically analyse behavioural patterns w.r.t. distinct user and performance groups.We utilise real-world data collected at the Lange Nacht der Musik (Long Night of Music) in Munich.

Analyzing Social Media Marketing in the High-End Fashion Industry Using Named Entity Recognition

Jorge Ale Chilet, Cuicui Chen and Yusan Lin

We study the marketing strategies of high-end fashion brands in social media. In particular, we focus on the informational content of brands’ posts in Instagram. Using Named Entity Recognition (NER) in Natural Language Processing (NLP), we develop a novel procedure to classify posts according to their information content. In addition, we apply NER to department store listings and expert runway reviews to obtain measures of brand leadership and brand similarity. Regression analyses show that, while follower brands respond to brand similarity and competitive pressure by relying on informational posts, the informational content of leaders presents a U-shaped relation with brand similarity.We interpret this finding using two theories from the marketing literature: the tradeoff between new and existing customers, and marketing life cycle of industries.

Analysis of the behavior of customers in the social networks using data mining techques

Leidys del Carmen Contreras Chinchilla and Kevin Andrey Rosales Ferreira

Companies today are developing business strategies taking into consideration behavior of their customers through social networks, which have allowed to extract large amounts of relevant data about users. This is why it has been necessary to apply data mining techniques to find patterns that describe the preferences of users in different contexts. This paper describes the results of using data mining techniques to analyze the behavior of customers of a fashion company in Instagram social network. The methodology used was CRISP-DM through which the descriptive models using the techniques of clustering and association rules were evaluated. The results shows that the proposed models can provide useful information to designing marketing strategies appropriate according to user preferences.

Evaluating the Impact of Social Media in Detecting Health-Violating Restaurants

Mikel Joaristi, Edoardo Serra and Francesca Spezzano

Nowadays, detecting health-violating restaurants is a serious problem due to the limited number of health inspectors in a city as compared to the number of restaurants. Rarely inspectors are helped by formal complains, but many complaints are reported as reviews on social media such as Yelp. In this paper we propose new predictors to detect healthviolating restaurants based on restaurant sub-area location, previous inspections history, Yelp reviews content, and Yelp users behavior. The resulting method outperforms past work, with a percentage of improvement in Cohen's kappa and Matthews correlation coefficient of at least %. In addition, we define a new method that directly evaluates the benefit of a classifier on the ability of an inspector in detecting health-violating restaurants. We show that our classification method really improves the ability of the inspector and outperforms previous solutions.

Agricultural activity shapes the mobility patterns in Senegal

S. Martin-Gutierrez, J. Borondo, A. J. Morales, J. C. Losada, A. M. Tarquis and R. M. Benito

The communication and migration patterns of a country are shaped by its socioeconomic processes. The economy of Senegal is predominantly rural, as agriculture employs over % of the labor force. In this work, we have used mobile phone records to explore the impact of agricultural activity on the mobility patterns of the inhabitants of Senegal. We have detected an increase in the migration flows throughout the country during the harvest season. At the same time, religious holidays also shape the mobility patterns of the Senegalese people, since, as in many cultures, they are related to climate seasons and agricultural activities. Hence, in the light of our results, agricultural activity and religious holidays are the primary drivers of mobility inside the country.

ASONAM - Multidisciplinary Track S2: Network Analysis and applications to information and Knowledge

Scheduled Seeding for Latent Viral Marketing

Alon Sela, Dmitri Goldenberg, Erez Shmueli and Irad Ben-Gal

One highly studied topic in the field of social networks is the search for influential nodes, that when seeded (i.e. infected intentionally), may infect a large portion of the network through a viral process. However, when it comes to the spread of new products, such viral processes are rather rare. Social influence is indeed an important factor when it comes to the act of adopting a new product. However, this influence is usually latent and does not trigger the purchase action by itself, it therefore requires an additional sales effort. We propose a model and a method that better fit the product adoption scenario. Our method allocates the seeding efforts not

only to precise nodes but also at precise points in time, such that the product adoption rate increases. By conducting a set of empirical simulations, we show that under realistic assumptions, our method improves the product adoption rate by %-%.

Evolution of MEDLINE bibliographic database: Preliminary Results

Andrej Kastrin, Thomas C. Rindflesch and Dimitar Hristovski

In this preliminary work we propose an approach to tracking network communities in time. We describe a methodology to study the dynamics and evolution of the MEDLINE bibliographic database using network-based analysis. We explore how the temporal characteristics of the network can be used to provide insight into the historical evolution of the broad field of biomedicine.

Impact of message sorting on access to novel information in networks

Benjamin D. Horne & Sibel Adal? and Kevin Chan

In social networks, individuals and systems work side by side. While individuals make decisions to filter or forward information, systems also prioritize and sort information to manage and assist individual information processing. It has long been argued that system level manipulations can reduce access of individuals to novel information. In this paper, we study how sorting of messages in one's inbox can help or hinder access of diverse information in the network through simulation of cognitively bounded actors. We show that first-in-first-out (FIFO) method of message sorting is ideal in bursty information arrival rates and in networks with lower diameter. Last-in-firstout (LIFO) method of message sorting is ideal for streaming information arrival, but leads to information overload in bursty scenarios by creating too many redundant copies of some of the information in the network. In short, the ideal message sorting method that enhances access to diverse information depends on the network type and information access patterns.

The Scientometrics of Successful Women in Science

Charisse Madlock-Brown and David Eichmann

This paper examines the effects of gender differences in collaboration on research outcomes. We analyzed network characteristics of seventeen medical research institutions that are Clinical and Translational Science Awardees (CTSA) to determine if network connectivity characteristics have the potential to help mitigate the performance gap between the sexes. We determined betweenness centrality to identity well connected researchers. Then we used clustering coefficient to determine how tightly connected their collaborators were with each other. We correlate these scores with productivity (number of total publications for each author), and h-index (the number of papers h for which an author has h citations). We also provide data on how network characteristics vary by role for each gender studied. Our results indicate that being well connected is more highly correlated with success for woman than men for most of the institutions we studied. We believe these results can be leveraged to improve success rates for women in the future

Selecting the cases that defined Europe: complementary metrics for a network analysis

Fabien Tarissan, Yannis Panagis and Urska Sadl

Do case citations reflect the "real" importance of individual judgments for the legal system concerned? This question has long been puzzling empirical legal scholars. Existing research typically studies case citation networks as a whole applying traditional network metrics stemming from graph theory. Those approaches are able to detect globally important cases, but since they do not take time explicitly into account, they cannot provide a comprehensive account of the dynamics behind the network structure and its evolution. In this paper we provide such a description, using two node importance metrics that take time into account to study

important cases in the Court of Justice of the European Union over time. We then compare cases deemed as important by the metrics, with a set of cases selected by the Court as the most important (landmark) cases. Our contribution is twofold. First, with regard to network science, we show that structural and time-related properties are complementary, and necessary to obtain a complete and nuanced picture of the citation network. Second, with regard to the case law of the Court, this study provides empirical evidence clarifying the motivation of the Court when selecting the landmark cases, revealing the importance of symbolic and historical cases in the selection. In addition, the temporal analysis sheds new light on the network properties specific to the landmark cases that distinguishes them from the rest of the cases. We validate our results by providing legal interpretations that sustain the highlights provided by the proposed network analysis.

Analysis and Visualization of a Literature-Mined Glaucoma Interaction Network

Maha Soliman, Olfa Nasraoui and Nigel G.F Cooper

Glaucoma is a silent eye disease that steals sight without warning and it is the second leading cause of blindness worldwide. It cannot be cured but it can be controlled. A gene network analysis to understand the cross talk between glaucoma networks of genes could contribute to better understanding and treatment of the disease. In this paper, we extend our previous text mining and network analysis research for extraction of a more comprehensive glaucoma interaction network from PubMed Central articles. After text mining of the PubMed Central literature to extract potential sentences bearing associations between glaucoma genes in the literature, the mined associations are used to construct a glaucoma gene interaction network and a network analysis is applied to understand the building structure of the network. The network analysis reveals a succinct network composed of five distinct clusters associated with seven benchmark glaucoma biomarkers. The knowledge that can be extracted from such a network could provide a useful summative knowledge base to complement other forms of clinical information related to this disease that affects many diagnosed individuals worldwide.

ASONAM - Multidisciplinary Track S3: Models and Methods for social media analysis

Model of computer architecture for online social network flexible data analysis: The case of Twitter data

Romain Giovannetti and Luigi Lancieri

Since several years, there is an increasing interest for new services based on the analysis of data coming from online social networks. Such services can, for example, provide the e-reputation of a product or a company, detect new trends in a commercial, social or political context, etc. The huge quantity of data is an opportunity in term of representativeness but is also difficult to manage. Within Twitter, for example, it appears that the huge stream of data is, most of the time, incompatible with a flexible analysis unless to have high computer resources. The only practical solution is often to observe in a static way a limited portion of a phenomenon in a limited time slot. This paper is devoted to the study of necessary conditions to provide an equilibrium between the computer architecture complexity and the analysis flexibility.

Finding Street Gang Members on Twitter

Lakshika Balasuriya, Sanjaya Wijeratne, Derek Doran and Amit Sheth

Most street gang members use Twitter to intimidate others, to present outrageous images and statements to the world, and to share recent illegal activities. Their tweets may thus be useful to law enforcement agencies to discover clues about recent crimes or to anticipate ones that may occur. Finding these posts, however, requires a method to discover gang member Twitter profiles. This is a challenging task since gang members represent a very small population of the million Twitter users. This paper studies the problem of automatically finding gang members on Twitter. It outlines a process to curate one of the largest sets of verifiable gang member profiles that have ever been studied. A review of these profiles establishes differences in the language, images, YouTube links, and emojis gang members use compared to the rest of the Twitter population. Features from this review are used to train a series of supervised classifiers. Our classifier achieves a promising F score with a low false positive rate.

Exploring public sentiments for livable places based on a crowd-calibrated sentiment analysis mechanism

Linlin You and Bige Tuncer

With the explosion of social networks, people more often share their opinions on-line, which provides a great opportunity to detect the public sentiment of a place in an automatic and timely way comparing to the conventional approaches, e.g., surveys, workshops and interviews. Even through the application of social sentiment analysis is widely discussed in many domains, e.g., politics, e-commerce, economy, and health and environment, to the best of our knowledge, no research has ever studied the effects of public sentiments of social networks in the domain of place design. In order to fill this vacancy, a sentiment analysis service, called geo-sentiment analysis service, is required, whose cores are) a social sentiment analysis engine, and) an intuitive and interactive visualization service. Thus, this paper firstly proposes CGSA: a Crowd-calibrated Geo-Sentiment Analysis mechanism, which can) start the sentiment analysis process based on the design of CTS (Compound Training Samples), and SSF (Social Sentiment Features),) perform three analyses, namely sentiment, clustering and time series analysis on geotagged social network messages, and) collect crowd-labelled data based on a crowdsourced calibration service to gradually improve the classification accuracy. As proved by two detailed analyses, SSF has the best accuracy in training sentiment classifiers, and the performance of the calibrated

classifier increases gradually and significantly from .% to .% in three calibration cycles. Moreover, as a part of a big project “Liveable Places”, “Sentiment in places” service with two visualization modes, namely D sentiment dashboard and D sentiment map, is implemented to support local authorities, urban designers and city planners better understand the effects of public sentiments regarding place (re)design in the testbed area: Jurong East, Singapore.

Modeling Twitter as Weighted Complex Networks Using Retweets

Muhammad Hrishiah, Maytham Safar and Khaled Mahdi

Social Networks are evolving rapidly to have a significant number of users using or joining their platforms daily, and a huge volume of information exchange and flow continuously. As human is the center of all information traffic, the information propagation became subjected to the relations and influences between the sources of these information and their followers or fans, these influences might be difficult to understand and model directly. While observing Twitter network, we emphasize on the importance of the retweet as a primary method for information propagation inside Twitter network. We present a detailed analysis for the behavior of the retweet functionality in the twitter platform, and then we show how our findings could shape the diffusion of the information in a network according to the relations and influences between users. Moreover, we propose measures to form a fully weighted directed graph where links and nodes have calculated values, and links are only assigned at the existence of communication transactions; the new measures can be used to predict influences and information diffusions. Finally, we demonstrate how using a retweet network for specific users, we can allocate key players and hidden identities who are potentially having the same ideology or maybe belong to the same group of interest.

Centrality measures in close group of adolescent females and their association with individual character strengths

Narotam Singh, Sheetal Varshney and Amita Kapoor

This paper amalgamates the field of positive psychology and social network analysis to explore what are the character strengths and virtues of individuals with high centrality measures within a close group of adolescent females. Research in the field of social network analysis in the last few decades has given a good understanding of different centrality measures. Today, we know what does an individual with high degree centrality, high betweenness centrality etc. signify in a group, this paper goes a step further, we attempt to find the correlation between individuals with high centrality measures and their character strengths. We analyze the friendship relationship of three different close groups of adolescent females studying in an undergraduate class. We construct the network of each class based on the friendship status and measure in-degree centrality, and betweenness centrality. The individuals with high centrality measures were then asked to undergo VIA character strength analysis. To investigate the correlation we quantified the character strength of the adolescent females with high centrality measures into numeric values using inverse transform and weighted sum technique. Our results show that an individual with highest in-degree centrality have kindness as one of her most prominent character strength, and individual with highest betweenness centrality have honesty, fairness and leadership as her prominent character strengths. Since an individual can be trained for particular character strengths, this correlation can help in training adolescent females (and may be extended to other groups) for specific role in an organization and society as a whole.

Unsupervised Models for Predicting Strategic Relations between Organizations

Shachi H Kumar, Jay Pujara, Lise Getoor, David Maresy, Dipak Guptaz and Ellen Riloff

Microblogging sites like Twitter provide a platform for sharing ideas and expressing opinions. The widespread popularity of these platforms and the complex social structure that arises within these communities provides a unique opportunity to understand the interactions between users. The political domain, especially in a multi-party system, presents compelling challenges, as political parties have different levels of alignment based on their political strategies. We use Twitter to understand the nuanced relationships between differing political entities in Latin America. Our model incorporates diverse signals from the content of tweets and social context from retweets, mentions and hashtag usage. Since direct communications between entities are relatively rare, we explore models based on the posts of users who interact with multiple political organizations. We present a quantitative and qualitative analysis of the results of models using different features, and demonstrate that a model capable of using sentiment strength, social context, and issue alignment has superior performance to less sophisticated baselines.

ASONAM - Multidisciplinary Track S4: Models and Methods for social media analysis

A Probalistic Approach to Automatically Extract New Words from Social Media

Geetika Sarna and MPS Bhatia

Social media is the collection of different social networks containing different type of information. The information may be in the form of text, video, audio and image. Also various categories of users, various types of communities are available on social network. This research reports on the extraction of new keywords from messages posted on social media which will be helpful in the identification of various communities, category of user and hidden pattern present in the social media. In this paper, we applied Probalistic approach to recognize the new keywords and assign the group accordingly. State-of-the- art studies performed detection on the basis of existing keywords but the proposed approach take decision based on the existing keywords and also on new keywords extracted from social media.

Multilevel Exploration in Twitter Social Stream

Luigi lancieri and Romain Giovanetti

This paper describes a methodology approach and a tool dedicated to the exploration of the twitter social stream by combining different contextual parameters such as time, keywords, gender or the opinion. The exploration can be made in two main modes depending on the fact that the phenomenon is either known or not. The first mode, similar to the use of Googlefight search engine, allows to compare the stream feedback for several groups of words. A typical example, that we will discuss, consists in evaluating trends in the domains of fashions or politic. The second mode consists in exploring the timeline of the social stream looking for unknown emerging events. This mode can be used to explore the past or to identify, in near real time, an event that will probably make the buzz.

Mimicry in Online Conversations: An Exploratory Study of Linguistic Analysis Techniques

Tom Carrick, Awais Rashid and Paul J Taylor

A number of computational techniques have been proposed that aim to detect mimicry in online conversations. In this paper, we investigate how well these reflect the prevailing cognitive science model, i.e. the Interactive Alignment Model. We evaluate Local Linguistic Alignment, word vectors, and Language Style Matching and show that these measures tend to show the features we expect to see in the IAM, but significantly fall short of the work of human classifiers on the same data set. This

reflects the need for substantial additional research on computational techniques to detect mimicry in online conversations. We suggest further work needed to measure these techniques and others more accurately.

Identifying Chinese Lexical Inference Using Probabilistic Soft Logic

Wei-Chung Wang and Lun-Wei Ku

Lexical inference problem is a significant component of some recent core AI and NLP research problems like machine reading and textual entailment. In this paper, we propose method utilizing the Probabilistic Soft Logic (PSL) model for Chinese lexical inference. The proposed PSL model not only can integrate two complementary traditional methods, i.e., the lexicalknowledge- based method and the distributional probabilistic method, but also can optimize the lexical inference network in a global view by the transitivity property of entailment relations. We build a large domain specific verb inference corpus containing , verb pairs with gold inference labels from math world problems. A five-folded experiment is performed. Results show that the proposed PSL model greatly outperforms our baseline .

Descriptive Group Detection in Two-mode Data Networks using Biclustering

Abdelilah Balamane, Rokia Missaoui, Leonard Kwuida and Jean Vaillancourt

The search for cohesive groups inside a social network is a topic commonly known as community detection and has attracted many researchers. However, the identification of groups with competitive features using blockmodeling, biclustering and structural or regular equivalences has benefited from a less important interest within the research community. In this paper we define a generic biclustering method called BiP that computes semantically meaningful coclusters (or biclusters) from a twomode data network. The method has the following features: (i) it allows the processing of adjacency matrices whose data type can be either binary, discrete or categorical without any need for prior data codification, and (ii) each generated block may either represent a cluster of objects with the properties they own (or not), or a juxtaposition of sub-groups with distinct profiles (and sometimes purely opposed ones) for a subset of attributes.

ASONAM - Multidisciplinary Track S5: New Algorithms for Data Mining and Media Analysis

ECO: Entity-level Captioning in Context

Hyunsouk Cho and Seung-won Hwang

Visual scene understanding has been one of the major goals of computer vision. However, existing work has focused on the object-level understanding, which limits the visual questions that can be answered. The goal of this paper is to invite collective efforts for entity-level understanding of images, by releasing ECO datasets and baselines for this task.

EXTRACT: New extraction algorithm of association rules from frequent itemsets

Ilhem Feddaoui, Fai?c?al Felhi and Jalel Akaichi

Stored data in database can hide some knowledge, which is interesting, useful to hidden knowledge discover. In this context, an algorithms number a frequent itemsets and association rules extraction were presented. Special feature of these algorithms is to generation a large number of rules, making their

exploitation a difficult task. In this paper we will introduce a new algorithm for association rules extraction. Proposed solution is based on two points, namely: frequent itemset extraction, and from these, it extracts association rules.

Detecting Community Patterns Capturing Exceptional Link Trails

Martin Aztmuller

We present a new method for detecting descriptive community patterns capturing exceptional (sequential) link trails. For that, we provide a novel problem formalization: We model sequential data as first-order Markov chain models, mapped to an attributed weighted network represented as a graph. Then, we detect subgraphs (communities) using exceptional model mining techniques:We target subsets of sequential transitions between nodes that are exceptional in that sense that they either conform strongly to a specific behavior or show significant deviations, estimated by a quality measure. In particular, such a community is described by a specific pattern composed of descriptive features (of the attributed graph) covering the respective community. We present a comprehensive modeling approach and discuss results of a case study in analyzing data from two real-world social networks.

A memory-efficient heuristic for maximum matching in scale-free networks

Upul Senanayake and Mahendra Piraveenan

The maximum matching problem has been extensively studied, and several algorithms have been proposed which can maximize the percentage of matching. Nevertheless, these algorithms are designed without consideration of the topology of the networks on which they are intended to be applied. However, recent research has shown that many distributed systems from social, technical and biological domains which can be represented as networks display the scale-free structure, which has well-known topological characteristics such as the power-law degree distribution. In this paper, we describe a simple iterative heuristic for maximum matchings in scale-free networks which takes advantage of such characteristics. We show that our heuristic is no worse than the best known version of the Blossom algorithm in terms of time-complexity, and much better than any version of Blossom in terms of memory usage (space-complexity) when applied to typical scale-free networks. The heuristic due to its simplicity and memory efficiency is a viable alternative to Blossom in most real world applications.

The Tale of Two Clocks

Zvi Lotker

The main question that this paper addresses is how to identify critical events in the evolution of a social network. The paper uses ideas from psychology about time perception. It is well known that time flows differently in different emotional situations. Equipped with this idea, this paper studies the relationship between two clocks. As opposed to standard synchronization, where everything is done in order to force clocks to agree on the time, the paper embraces the discrepancy between the clocks. This paper presents a standard model where two natural clocks exists simultaneously: the event clock Ce and the weighted clock Cw. As the paper shows, using the drift between those two clocks is useful to understand the dynamics in social networks. The main claim is that the drift between different clocks points to a critical event in the evolution of the social network, similar to time perception in psychology. In order to demonstrate this claim, plays by William Shakespeare were used, and from them two clocks were created: the ”word time”, which is the weighted clock, and the ”response time”, which is the event clock. The paper will introduce the concept of a single clock drift. A play can have many, or a single

clock drift events. It is shown that in the single clock drift plays, the beginning of the drift points to a critical event in the play. The results are compared with the ”standard common” opinion.

ASONAM - Multidisciplinary Track S6: Social Media Analysis and Political Issues

Toward understanding how users respond to rumors in social media

Anh Dang, Mike Smit, Abidalrahman Moh!&d, Rosane Minghim and Evangelos Milios

As the spread of rumours has been increasing every day in online social networks (OSNs), it is important to analyze and understand this phenomenon. Damage caused by the spread of rumours is difficult to handle without a full understanding of the dynamics behind it. One of the central steps of understanding rumour spread is to analyze who spread rumours online, why, and how. In this research, we focus on the steps who and why by describing, implementing, and evaluating an approach that studies whether or not a group of users is actively involved in rumour discussions, and assesses rumour-spreading personality types in OSNs. We implement this general approach using Reddit data, and demonstrate its use by determining which users engage with a recurring rumour, and analyzing their comments using qualitative methods. We find that we can reliably classify users into one of three categories: () “Generally support a false rumour”, () “Generally refute a false rumour”, or () “Generally joke about a false rumour”. Combining text mining techniques, such as text classification, sentiment analysis, and social network analysis, we aim to identify and classify those rumour-spreading user categories automatically and provide a more holistic view of rumour spread in OSNs.

Analyzing the Usage of Social Media During Spanish Presidential Electoral Campaigns

J. Borondo, A. J. Morales, J. C. Losada and R. M. Benito

The large amount of user generated data that Online Social Networks produce has remarkably drawn the attention for researchers on human behavior in the recent years. In this work, we use temporal series and complex network analysis to unveil the users’ behavioral patterns during the Spanish presidential electoral campaigns in Twitter. We introduce a new measure to study political sentiment in Twitter, which we call the relative support. We have also characterized user behavior by analyzing the structural and dynamical patterns of the complex networks emergent from the mention and retweet networks. Our results suggest that the collective attention is driven by a very small fraction of users. Furthermore, we have analyzed the interactions taking place among politicians, observing a lack of debate. Moreover, we characterize the users and politicians’ interactions and propose a model to simulate their behavior.

The social media election agenda: Issue salience on Twitter during the European and Swedish elections

Linn Sandberg, Shatha Jaradat and Nima Dokoohaki

the role of issues in electoral preference formation has long been an established key factor and what voters consider the most important problem i.e. issue salience is essential for party choice. Political issues (and their salience to the electorate) also play an important role in parties’ tactical campaign strategies. This study examines to what extent social media possibly can contribute in shaping the issue agenda regarding the political parties. The issue agenda on Twitter is likely to have its own characteristics and dynamics, shaped by the technical peculiarities, users and the new campaigning possibilities that social media offers. This study will identify what issues are salient in the online discussions in conjunction with the European election and Swedish national election . The distribution

of issue attention divided to the various parties on social media is analyzed in light of the issue agenda set forth by the voters for the different elections.

An Analysis of Sentiments on Facebook during the U.S. Presidential Election

Saud Alashri, Srinivasa Srivatsav Kandala, Vikash Bajaj, Roopek Ravi, Kendra L. Smith and Kevin C. Desouza

Social networking sites (SNS), such as Facebook and Twitter, are important spaces for political engagement. SNS have become common elements in political participation, campaigns, and elections. However, little is known about the dynamics between candidate posts and commentator sentiment in response to those posts on SNS. This study enriches computational political science by studying the U.S. elections and how candidates and commentators engage on Facebook. This paper also examines how online activity might be connected to offline activity and vice versa. We extracted , Facebook posts by five presidential candidates (Hillary Clinton, Donald Trump, Bernie Sanders, Ted Cruz, and John Kasich) from their official Facebook pages and ,, comments on those posts. We employed topic modeling, sentiment analysis, and trends detection using wavelet transforms to discover topics, trends, and reactions. Our findings suggest that Republican candidates are more likely to share information on controversial events that have taken place during the election cycle, while Democratic candidates focus on social policy issues. As expected, commentators on Republican candidate pages express negative sentiments toward current public policies as they seldom support decisions made by the Obama administration, while commentators on democratic candidate pages are more likely to express support for continuation or advancement of existing policies. However, the significance (strong/weak) and nature (positive/negative) of sentiments varied between candidates within political parties based on perceived credibility of the candidate's degree of credibility on a given issue. Additionally, we explored correlation between online trends of comments/sentiment and offline events. When analyzing the trend patterns, we found that changes in online trends are driven by three factors:) popular post,) offline debates, and) candidates dropping out of the race.

Social Media, Spillover, and Saudi Arabian Women!&s right to drive movements: Analyzing interconnected online collective actions

Serpil Tokdemir, Nitin Agarwal and Rolf T. Wigand.

The advent of modern forms of information and communication technologies (ICTs), such as social media, has modified the ways people communicate and enables an inimitable way of connectivity promoting the advancement and diffusion of information. Given that, individuals within social movements bringing influence from other movements makes the study of spillover within social movements an important concentration for sociologists and others studying collective action (CA). This research explores the role of social movement spillover in investigating and sustaining the Women's Right to Drive movement in the Kingdom of Saudi Arabia. We benefit from various models of established collective action theories developed in the pre-Internet era, and re-evaluate traditional theories of spillover within the modern ICT landscape by utilizing existing collective action theories/approaches and novel and innovative computational analytical tools. The findings of this study are conceptualized to shed new insights on information diffusion, mutual influence, role distribution analysis of activists/supporters across movements and provide a deeper understanding of interconnected social movements and social movement spillover. Applying computational metrics to measure the strength of spillover and its effects on complex social processes enable novel model development to help advance the understanding of interconnected collective actions conducted through modern social and information systems.

FOSINT-SI Session 1

Spam Detection of Twitter Traffic: A Framework based on Random Forests and non-uniform feature sampling

Claudia Meda, Edoardo Ragusa, Christian Gianoglio, Rodolfo Zunino, Augusto Ottaviano, Eugenio Scillia and Roberto Surlinelli

Law Enforcement Agencies cover a crucial role in the analysis of open data and need effective techniques to filter troublesome information. In a real scenario, Law Enforcement Agencies analyze Social Networks, i.e. Twitter, monitoring events and profiling accounts. Unfortunately, between the huge amount of internet users, there are people that use microblogs for harassing other people or spreading malicious contents. Users' classification and spammers' identification is a useful technique for relieve Twitter traffic from uninformative content. This work proposes a framework that exploits a non-uniform feature sampling inside a gray box Machine Learning System, using a variant of the Random Forests Algorithm to identify spammers inside Twitter traffic. Experiments are made on a popular Twitter dataset and on a new dataset of Twitter users. The new provided Twitter dataset is made up of users labeled as spammers or legitimate users, described by features. Experimental results demonstrate the effectiveness of enriched feature sampling method.

Virtual Indicators of Sex Trafficking to Identify Potential Victims in Online Advertisements

Michelle Ibanez and Rich Gazan

A content analysis of online sex worker advertisements suggests specific terms, sources and patterns of behavior that may help identify potential sex trafficked victims within these virtual environments. While some ads are posted by independent sex workers, others may have been posted by traffickers or pimps, advertising the women they have under their control. A total of ads from the Backpage "escort" services site were harvested and analyzed for information elements including location, age, name, area code, ethnicity and controlled/restricted movement. The results suggest that % of the ads contained one or more of the primary sex trafficking indicators, and % of the ads contained three or more indicators. The results suggest that some physical indicators should be adapted to the virtual environment, and that the presence of multiple indicators can be used to prioritize certain ads for further investigation by law enforcement. It is hoped that these indicators can inform the design of future systems to flag high-risk advertisements, and help identify and disrupt this covert network activity.

Investigative Simulation: Towards Utilizing Graph Pattern Matching for Investigative Search

Benjamin Hung and Anura Jayasumana

This paper proposes the use of graph pattern matching for investigative graph search, which is the process of searching for and prioritizing persons of interest who may exhibit part or all of a pattern of suspicious behaviors or connections. While there are a variety of applications, our principal motivation is to aid law enforcement in the detection of homegrown violent extremists. We introduce investigative simulation, which consists of several necessary extensions to the existing dual simulation graph pattern matching scheme in order to make it appropriate for intelligence analysts and law enforcement officials. Specifically, we impose a categorical label structure on nodes consistent with the nature of indicators in investigations, as well as prune or complete search results to ensure sensibility and usefulness of partial matches to analysts. Lastly, we introduce a natural top-k ranking scheme that can help analysts prioritize investigative efforts. We demonstrate performance of investigative simulation on a real-world large dataset.

The Rise & Fall of #NoBackDoor on Twitter: the Apple vs. FBI Case

Samer Al-Khateeb and Nitin Agarwal

In addition to using social media to connect with others worldwide, many people nowadays get their news about different national or international events such as natural disasters, crises, political elections, conflicts etc. via social media. This evolution in the usage of social media has not only led to the generation of massive amounts of data but also various information consumption behaviors. In this study, we developed a framework that can be used to monitor/understand, analyze, and visualize in real time how people consume information and react to events. Following the case study of Apple, Inc. vs. FBI, we tracked the usage of the #NoBackDoor on Twitter in real time and were able to understand what people are thinking about the case and who are the actors involved in this network. The framework can be applied to study other events and provide a deeper understanding of how public sentiments evolve during an event, whether it is a crisis or major news event.

FOSINT-SI Session 2

Argumentation Models for Cyber Attribution

Eric Nunes, Paulo Shakarian, Gerardo Simari and Andrew Ruef

A major challenge in cyber-threat analysis is combining information from different sources to find the person or the group responsible for the cyber-attack. It is one of the most important technical and policy challenges in cyber-security. The lack of ground truth for an individual responsible for an attack has limited previous studies. In this paper, we take a first step towards overcoming this limitation by building a dataset from the capture-the-flag event held at DEFCON, and propose an argumentation model based on a formal reasoning framework called DeLP (Defeasible Logic Programming) designed to aid an analyst in attributing a cyber-attack. We build models from latent variables to reduce the search space of culprits (attackers), and show that this reduction significantly improves the performance of classification-based approaches from % to % in identifying the attacker.

FOSINT-SI Session 3

Graph Analytics for Healthcare Fraud Risk Estimation

L. Karl Branting, Flo Reeder, Jeff Gold and Timothy Champney

This paper presents a novel approach to estimating healthcare fraud (HCF) risk that applies network algorithms to graphs derived from open source datasets. One group of algorithms calculates behavioral similarity to known fraudulent and non-fraudulent healthcare providers with respect to measurable healthcare activities, such as medical procedures and drug prescriptions. Another set of algorithms estimates propagation of risk from fraudulent healthcare providers through geospatial collocation, i.e., shared practice locations or other addresses. The algorithms were evaluated with respect to their ability to predict a provider's presence on the Office of the Inspector General's list of providers excluded from participation in Medicare and other Federal healthcare programs (exclusion). In an empirical evaluation, a combination of features achieved an f-score of . and a ROC area of . in exclusion prediction. An ablation analysis showed that most of this predictive accuracy was the result of features that measure risk propagation through geospatial collocation.

Retweet Prediction Considering User's Difference as an Author and Retweeter

Syeda Firdaus, Chen Ding and Alireza Sadeghian

Social network is a hot topic of interest for the researchers in the field of computer science in recent years. The vast amount of data generated by these social networks play a very important role in information diffusion. Social network data are generated by its users. So, user's behavior and activities are being investigated by the researchers to get a logical view of social network platform. This research proposed a novel retweet prediction model which considers difference in user's behavior as an author (as reflected in the tweets) and a retweeter (as reflected in the retweets) and do the prediction accordingly. The proposed retweet prediction strategy taking this difference into consideration, gave better prediction accuracy than the conventional strategy. The findings of this research explains that in social networks, some users show different behavior indifferent roles and these differences may have impact on future research.

CyberTwitter: Using Twitter to generate alerts for Cybersecurity Threats and Vulnerabilities

Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi and Tim Finin

In order to secure vital personal and organizational system we require timely intelligence on cybersecurity threats and vulnerabilities. Intelligence about these threats is generally available in both overt and covert sources like the National Vulnerability Database, CERT alerts, blog posts, social media, and dark web resources. Intelligence updates about cybersecurity can be viewed as temporal events that a security analyst must keep up with so as to secure a computer system. We describe CyberTwitter, a system to discover and analyze cybersecurity intelligence on Twitter and serve as a OSINT (Open–source intelligence) source. We analyze real time information updates, in form of tweets, to extract intelligence about various possible threats. We use the Semantic Web RDF to represent the intelligence gathered and SWRL rules to reason over extracted intelligence to issue alerts for security analysts.

Hidden Social Networks Analysis by Semantic Mining of Noisy Corpora

Christophe Thovex

The present work proposes a paradigm for the analysis of social networks hidden within incomplete data models, based on the semantic mining of noisy corpora. A proof of concept is implemented and experimented on a partial database resulting from the capture of short text messages in line with the international project ‘Smsscience’.

FOSINT-SI Session 4

Detecting Covert Sex Trafficking Networks in Virtual Markets

Michelle Ibanez and Dan Suthers

Covert sex trafficking networks are increasingly using information and communication technologies (ICTs) to extend their operations. There is a need for systematic research and methods for the study of technology facilitated sex trafficking. This study examined how publicly available information can be used to uncover covert sex trafficking networks. The intent was to transform the types of data available in online advertisements into meaningful information that can be used to disrupt this activity. Content analysis was used to identify important data fields in online escort advertisement that presented virtual indicators of sex trafficking, and social network analysis methods were applied to identify provider networks.

Temporal Analysis of Dark Web Forum Users

Andrew Park, Brian Beck, Darrick Fletcher, Patrick Lam and Herbert H. Tsang

Extremist groups have turned to the Internet and social media sites as a means of sharing information amongst one another. This research study analyzes forum posts and finds people who show radical tendencies through the use of natural language processing and sentiment analysis. The forum data being used are from six Islamic forums on the Dark Web which are made available for security research. This research project uses a POS tagger to isolate keywords and nouns that can be utilized with the sentiment analysis program. Then the sentiment analysis program determines the polarity of the post. The post is scored as either positive or negative. These scores are then divided into monthly radical scores for each user. Once these time clusters are mapped, the change in opinions of the users over time may be interpreted as rising or falling levels of radicalism. Each user is then compared on a timeline to other radical users and events to determine possible connections or relationships. The ability to analyze a forum for an overall change in attitude can be an indicator of unrest and possible radical actions or terrorism.

Cyberbullying Detection Using Probabilistic Socio-Textual Information Fusion

Vivek Singh, Qianjia Huang and Pradeep K. Atrey

Cyberbullying is an important socio-technical challenge in Online Social Networks (OSN).With the growth trends of heterogeneous data in OSN, better network characterization, and textual feature sophistication, recent efforts have realized the value of looking at heterogeneous modes of information including textual features, social features, and image-based features for better cyberbullying detection. These approaches, however, still use these features either individually or combine them 'naively' without considering the different confidence levels associated with each feature or the interdependencies between features. We propose a novel probabilistic information fusion framework that utilizes confidence score and interdependencies associated with different social and textual features and uses those to build better predictors for cyberbullying. The performance of the proposed approach was compared to a recent approach in literature which used a similar dataset and features and the proposed approach resulted in significant improvements in terms of cyberbullying detection.

Understanding Alliance and Opposition Among Violent Groups

Quan Zheng and David Skillicorn

In many parts of the world there are complex, violent interactions among groups with widely varying agendas. Situational awareness is difficult because there is rarely a clear distinction between good and bad actors, and there are constantly shifting alliances and oppositions between groups. This makes it difficult for analysts to understand the ecosystem of a country and region; still more to conceive of helpful interventions. We show how to use a newly developed spectral graph embedding technique that allows social networks with edge weights that are positive (alliance) and negative (opposition) to be modelled. We show the practical application by applying the technique to countries in North-West Africa, where civil wars are commonplace and complex islamist insurgencies have been active in the past few decades. A sense of the differences among these countries becomes visible, as well as a picture of the interactions among the key groups within each country.

Detecting Sex Trafficking Circuits in the U.S. Through Analysis of Online Escort Advertisements

Michelle Ibanez and Rich Gazan

The use of social network analysis is used to identify possible victims of sex trafficking and the observation of domestic trafficking flows across the U.S. A novel approach to the analysis is presented with the understanding further development is needed. Online escort advertisements were analyzed for virtual indicators of sex trafficking and to identify patterns of activity that may be associated with criminal networks. The results suggest that % of the sample contained indicators of movement. Social network analysis methods were applied to identify movement trends across the US. The use of SNA methods allowed prominent hubs and circuits of this activity to be observed, by providing a tool to uncover covert network structures and activity, yielding a method to capture movement trends of potential trafficked persons. The integration of geospatial data allowed maps to be created in order to visualize movement patterns.

FAB : Session 1 - Recommendation Systems

FriendRank: A Personalized Approach for Tweets Ranking in Social Networks

Min Li, Linfeng Luo, Lin Miao, Yibo Xue, Zhiyun Zhao and Zhenyu Wang

The thousands of streaming data overwhelmingly provide for Internet users on Twitter every day, especially for those Twitter users with many friends. However, the useful tweets that users are really interested in personally could be covered by massive other uninformative and uninteresting information. Therefore, how to bring immediately the interesting tweets for users is always a challenging issue. In this paper, we consider the user friendships in detail and build an effective and practical model to calculate the friendships among users. Certainly, we also take user interests to tweets into account. We then propose a personalized approach for tweets ranking, which focus on the user friendships and the personal interests to tweets. The experimental results demonstrate that our proposed method greatly outperforms several baselines and the

Personalized Recommendation for New Questions in Community Question Answering

Lin Wang, Bin Wu, Juan Yang and Shuang Peng

Community question answering(CQA) websites such as Yahoo! Answers and Stack Overflow provide a new way of asking and answering questions which are not well served by general web search engines. Due to the huge volume and everincreasing number of questions, not all new questions can get fully answered in required time. Therefore, it is of great significance to design some effective strategies of recommending experts for new questions. In this paper, we propose a novel personalized recommendation method for routing new questions to a group of experts. Different from prior work which only considers the topic modeling or the link structure, we aim at recommending new questions to more appropriate experts by considering both of these two factors. Moreover, we design a new strategy of network construction with the personalization fully considered. The comparison experiments are conducted with Stack Overflow data and the experimental results demonstrate that the proposed method improves the recommendation performance over other methods in expert recommendation.

Time Preference aware Dynamic Recommendation Enhanced with Location, Social Network and Temporal Information

Makbule Gulcin Ozsoy, Faruk Polat and Reda Alhajj

Social networks and location based social networks have many active users who provide various kind of data, such as where they have been, who their friends are, which items they like more, when they go to a venue. Location, social network and temporal information provided by them can be used by recommendation systems to give more accurate suggestions. Also, recommendation systems can

provide dynamic recommendations based on the users' preferences, such that they can give different recommendations for different hours of the day or different days of the week. In this paper, we propose a recommendation system which considers the users' temporal preference to give dynamic recommendation. The recommendation method uses multi-objective optimization approach and gives point of interest (POI) recommendation using several different criteria, namely past check-in locations, hometown of users, time of check-ins, friendship and influence among users.

An Intelligent Method for Optimization of Tariffs in GSM Networks

Buket Kaya Sefa Sahin Koc

GSM is a mobile technology that allows people to communicate with one another. The technology enables people to call others over the phone with a GSM number and a certain tariff for communication. An interpersonal GSM network is established by means of calls and text messages. This paper proposes an approach to recommend optimal tariffs to GSM users to maximize the total utility of individuals in the GSM network. However, finding optimal tariffs for very large GSM networks is computationally intractable by standard methods. For this purpose, we present a novel multi-agent based algorithm to recommend the most appropriate tariffs to a specific group of GSM users. The results of the experiment on a synthetic GSM network comprised of users with various tariffs suggest that the method is practical and able to yield accurate results.

Big Data Mining of Social Networks for Friend Recommendation

Fan Jiang, Carson K. Leung and Adam G. M. Pazdor

In the current era of big data, high volumes of valuable data can be easily collected and generated. Social networks are examples of generating sources of these big data. Users in these social networks are often linked by some interdependency such as friendship. As these big social networks keep growing, there are situations in which an individual user wants to find popular groups of friends so that he can recommend the same groups to other users. In this paper, we present a big data analytic solution that uses the MapReduce model in mining these big social networks for discovering groups of frequently connected users for friend recommendation. Evaluation results show the efficiency and practicality of our data analytic solution in mining big social networks, discovering popular users, and recommending friends.

FAB : Session 2 - Pattern Detection

Mining 'Following' Patterns from Big Sparse Social Networks

Carson K. Leung, Edson M. Dela Cruz, Trevor L. Cook and Fan Jiang

In the current era of big data, high volumes of valuable data can be easily collected and generated. Social networks are examples of generating sources of these big data. Users (or social entities) in these social networks are often linked by some interdependency such as friendship or 'following' relationships. As these big social networks keep growing, there are situations in which an individual user (or business) wants to find those frequently followed groups of social entities so that he can follow the same groups. Discovery of these frequently followed groups can be challenging because the social networks are usually big (with lots of users/social entities) but sparse (with most users only know some but not all users/social entities in a social network). In this paper, we present a data analytic solution that uses a compression model in mining these big but sparse social networks for discovering groups of

frequently followed social entities. Evaluation results show the efficiency and practicality of our data analytic solution in discovering ‘following’ patterns from social networks.

Frequent and Non-Frequent Pattern Detection in Big Data Streams: An Experimental Simulation in Trillion Data Points

Konstantinos Xylogiannopoulos, Panagiotis Karampelas and Reda Alhajj

Big data streaming analysis nowadays has become one of the most important topic in the list of data analysts since enormous amount of data are produced daily by the numerous smart devices. The analysis of such data is very important and the detection of frequent or even non-frequent patterns can be critical for many aspects of our lives. In the current paper, we propose a new methodology based on our previous work regarding the detection of all repeated patterns in a string in order to analyze a very big data stream with Trillion digits, composed from thousand subsequences of billion digits each one. More specifically, using the novel data structure, LERP Reduced Suffix Array, and the innovative ARPaD algorithm which allows the detection of all repeated patterns in a string we managed to analyze each one of the billion data points, using computers with standard hardware configuration, in minutes which outperforms to the best of our knowledge any other existing methodology, which is equivalent to data point generation every microseconds.

Mining Quad Closure Patterns in Instagram

Ahmet An?l Mungen Mehmet Kaya

Quad Closure is a group of four people who are connected with each other. In this paper, we propose a new group recognition method for Instagram which are based on triadic closure method to determine groups on dynamic social networks (e.g likes and comments) between users as named Quad Closure. Social networks are not easily classified because of their complexity. We study how an open quad closure becomes close quad and the possibility of finding this process. There are a wide variety of factors having effects on this process. These factors include users' background information like active using time piece, photos tag frequency and sentimental analyze on comments. We try to create a unique model to predict formation of quad closure with all these factors and we share our experimental results on data taken from Instagram and show the success of our method on quad closure patterns.

Co-Clustering Signed -Partite Graphs

Sefa Sahin Koc, Ismail Hakk? Toroslu and Hasan Davulcu

In this paper, we propose a new algorithm, called STRICLUSTER, to find tri-clusters from signed -partite graphs. The dataset contains three different types of nodes. Hyperedges connecting three nodes from three different partitions represent either positive or negative relations among those nodes. The aim of our algorithm is to find clusters with strong positive relations among its nodes. Moreover, negative relations up to a certain threshold is also allowed. Also, the clusters can have no overlapping hyperedges. We show the effectiveness of our algorithm via several experiments.

Classification of HIV data By Constructing A Social Network with Frequent Itemsets

Yunuscan Kocak Tansel Ozyer Reda Alhajj

Acquired immune deficiency syndrome (AIDS) is the last and the most life-threatening phase of Human Immunodeficiency Virus (HIV) disease. HIV attacks and heavily affects the immune system of the body which remains unable to resist the disease. HIV uses white blood cells to replicate itself and spreads everywhere in the body. The lifecycle of HIV disease, especially the replication stage must be

prominently understood in order to develop effective drugs for treatment. HIV- protease enzyme is in charge of cleaving an amino acid octamer into peptides which are used to create proteins by virus. It should be scrutinized properly since it is a potential target to tightly bind drugs to protease for blocking the virus action at an early stage before cell infection. It is very critical to induce a model and predict cleavage of HIV- protease on octamers. Several machine learning approaches have been applied for predicting and profiling cleavage rules. However, we propose a novel general approach that can also be applied on different domains. It basically utilizes social network analysis and data mining techniques for classification. This method yet presents promising results that are comparable with existing machine learning methods, besides it gives the opportunity to validate the results obtained by using other techniques from social network analysis perspective. We have used the HIV- protease cleavage data set from UCI machine learning repository and demonstrated the effectiveness of our proposed method by comparing it with decision tree, Naive-Bayes and k-nearest neighbor methods

FAB : Session 3 - Machine Learning Methods

A New Topological Metric for Link Prediction in Directed, Weighted and Temporal Networks

Ertan Butun, Mehmet Kaya and Reda Alhajj

One of the most interesting tasks in social network analysis is link prediction. There are a lot of studies dealing with link prediction task in the literature. In recent years, there is an increasing on link prediction methods trying to model network as more close to real networks such as heterogeneous, temporal and directed network models to gain better link prediction performance. Many of the existing link prediction methods don’t take into account links directions in directed networks. In this paper we propose a new neighbor and graph pattern based topological metric considering direction of links for link prediction. The proposed metric also takes into account temporal and weighted information, which are useful to increase link prediction performance. Accuracy of the proposed metric is evaluated by comparison with multiple baseline metrics from literature in supervised learning methods. Experimental results demonstrate that the proposed metric improves remarkably the accuracy of link prediction.

Spectral Graph-Based Semi-supervised Learning for Imbalanced Classes

Quan Zheng and David Skillicorn

Semi-supervised learning makes the realistic assumptions that labelled data is typically rare, and that unlabelled data that are similar are likely to belong to the same class. Unlabelled data are assigned the labels associated with their “most similar” labelled neighbors. For graph-based semi-supervised learning, “most similar’ is defined by weighted multipath path length in a graph. When classes are of different sizes, or the number of labelled nodes per class is not the same across classes, the performance of existing graph-based algorithms degrades sharply. We develop a new algorithm that creates representative nodes for each class, connects them to the labelled nodes of that class, adds negative edges between them, embeds the resulting graph using a signed graph Laplacian technique, and then predicts the unlabelled nodes using distance-based techniques in the geometry of the embedding. Its performance matches current algorithms for balanced datasets, but is much better for datasets where the classes, or the number of labelled records, differ in size. Keywords: spectral graph embedding, signed graphs, semisupervised learning, Laplacians

Predicting Good Fit Students by Correlating Relevant Personality Traits with Academic/Career Data

Muhammad Fahim Uddin Jeongkyu Lee

This paper discusses part of the main work in field of data science, mining and analytics. Family of algorithms is developed to predict the educational relevance of individuals' talents through lens of personality features (unstructured and semi-structured) and academic/career data. This paper presents progress of results in Good Fit Students (GFS) algorithms and math construct. This work addresses the problems of poor academic performances, low retention rates, drop outs, school transfers, costly readmissions, poor job performances, early job transfers and inefficient utilization/consideration of natural talents. GFS builds a framework and algorithms by correlating and blending social networking personality traits data with academic and career data. The results are promising at this stage of research and show improved predictions and relevant probabilities. Future work is focused on improving the results with more data and adding few more algorithms to the main research/framework.

Examining Place Categories for Link Prediction in Location Based Social Networks

Ahmet Engin Bayrak and Faruk Polat

The day mankind met with smart-phones, a new era started. Since then, daily mobile internet usage rates are increasing everyday and people have developed new habits like frequently sharing information (photo, video, location, etc.) on online social networks. Location Based Social Networks (LBSNs) are the platforms that empowers users to share place/location information with friends. As all other social networks, LBSNs aim to acquire more users with a smart friend recommendation. Solution for smart friend recommendation problem is studied under link prediction field by researchers. Check-in information is the main data for link prediction in LBSNs. Data extracted from check-in information plays vital role for predictor performance. In this study, we attempt to make use of detailed analysis of place category in order to exploit possible information gain enhancements through such semantic information. We proposed two new feature groups; Common Place Check-in Count Product Sum and Common Category Check-in Count Sum Product. For any link candidate pair; those features are calculated for each category. Use of new features improved the link prediction performance for multiple data subsets.

FAB : Session 4 - Social Network Applications

Spiral of Silence in Social Networks: A Data-driven Approach

Linfeng Luo, Min Li, Qing Wang, Yibo Xue, Chunyang Liu and Zhenyu Wang

Although the spiral of silence theory has been studied thoroughly in the traditional dissemination field, to our best knowledge, no one has clearly verified the applicability of the spiral of silence theory in social networks based on the real information propagation datasets. In this paper, we focus on the disparity between majority and minority opinions, we verify the applicability of the spiral of silence theory in social networks by taking into account factors, including the propagation width, the propagation depth, the message sentiment and the modularity through a large amount of data-driven experiments based on the real-world information propagation datasets which collected on Sina Weibo. We also investigate the applicability of tweets with different categories, our data-driven experimental results show that the spiral of silence theory is still applicable in social networks but different tweets with different categories have different applicability of the spiral of silence theory.

On data collection, graph construction, and sampling in Twitter

Jeremy D. Wendt, Randy Wells, Richard V. Field and Jr. Sucheta Soundarajan

We present a detailed study on data collection, graph construction, and sampling in Twitter. We observe that sampling on semantic graphs (i.e., graphs with multiple edge types) presents fundamentally distinct challenges from sampling on traditional graphs. The purpose of our work is to present new challenges and initial solutions for sampling semantic graphs. Novel elements of our work include the following: () We provide a thorough discussion of problems encountered with na¨ıve breadth-first search on semantic graphs. We argue that common sampling methods such as breadth-first search face specific challenges on semantic graphs that are not encountered on graphs with homogeneous edge types. () We present two competing methods for creating semantic graphs from data collects, corresponding to the interactions between sampling of different edge types. () We discuss new metrics specific to graphs with multiple edge types, and discuss the effect of the sampling method on these metrics. () We discuss issues and potential solutions pertaining to sampling semantic graphs.

An Experimental Evaluation of Giraph and GraphChi

Junnan Lu and Alex Thomo

We focus on the vertex-centric (VC) model introduced in Pregel, a Google system for distributed graph processing. In particular, we consider two popular implementations of the VC model: Apache Giraph and GraphChi. The first is a VC system for cluster computing, while the second is a VC system for a single PC. Apache Giraph became very popular after careful engineering by Facebook researchers in to scale the computation of PageRank to a trillion-edge graph of user interactions using machines. On the other hand, GraphChi became popular, around the same time in , as it made possible to perform intensive graph computations in a single PC, in just under minutes, whereas the distributed systems were taking minutes using a cluster of about , computers (as reported also by MIT Technology Review). Since then, new versions of Apache Giraph and GraphChi have been released, where new ideas and optimizations have been implemented. Therefore, it is time to validate again the claims made four years ago. In this work, we embark in this validation. We consider three cornerstone graph problems: computing PageRank, shortest-paths, and weakly-connected-components. Based on current experiments, we conclude that in the present, even for a moderate number of simple machines, Apache Giraph outperforms GraphChi for all the algorithms and datasets tested. This is in contrast to the claims of the GraphChi authors in .

A Privacy Weaving Pipeline for Open Big Data

Yuan-Chih Yu and Dwen-Ren Tsai

The power of big data gives us an unprecedented chance to understand, analyze, and recreate the world, while open data ensures that power be shared and widely exploited. Open and big data has become the emerging topics for researchers and governments. Thus, the related privacy issues also become an emerging urgent problem. In this work, we propose a conceptual framework of privacy weaving pipeline dedicated for producing open and big data while preserving privacy. Within the processing pipeline, each step of the process flow considers the privacy assurance to manipulate datasets. However, the complexity of process flow is the same as normal data pipeline. The experimental prototype confirms the feasibility of framework design. We hope this work will facilitate the development of open and big data industry.

HIBIB -- Session 1

Study of transductive learning and unsupervised feature construction methods for biological sequence classification

Ana Stanescu, Karthik Tangirala and Doina Caragea

Next Generation Sequencing (NGS) technologies have led to fast and inexpensive production of large amounts of biological sequence data, including nucleotide sequences and derived protein sequences. These fast-increasing volumes of data pose challenges to computational methods for annotation. Machine learning approaches, primarily supervised algorithms, have been widely used to assist with classification tasks in bioinformatics. However, supervised algorithms rely on large amounts of labeled data in order to produce quality predictors. Oftentimes, labeled data is difficult and expensive to acquire in sufficiently large quantities. When only limited amounts of labeled data but considerably larger amounts of unlabeled data are available for a specific annotation problem, semi-supervised learning approaches represent a cost-effective alternative. In this work, we focus on a special case of semi-supervised learning, namely transductive learning, in which the algorithm has access during the training phase to the instances that need to be labeled. Transduction is particularly suitable for biological sequence classification, where the goal is generally to label a given set of unlabeled instances. However, a challenge that needs to be addressed in this context consists of identification of compact sets of informative features. Given the lack of labeled data, standard supervised feature selection methods may result in unreliable features. Therefore, we study recently proposed unsupervised feature construction approaches together with transductive learning. Experimental results on two classification problems, namely cassette exon identification and protein localization, show that the unsupervised features result in better performance than the supervised features.

Understanding Crohn!&s disease patients reaction to Infliximab from Facebook: a medical perspective

Marco Roccetti, Paola Salomoni, Catia Prandi, Gustavo Marfia, Marco Montagnani and Linda Gningaye

This paper completes an analysis started with two previous contributions, which have first identified Facebook as the most interesting social media for Crohn’s disease patients and then Infliximab as the most discussed, both positively and negatively, treatment. In particular, in our second contribution, we concentrated on the satisfaction that emerged from online posts regarding Infliximab and compared it to the body of pertaining medical literature. We here finish off such work, comparing the satisfaction recorded by automatic means through the use of sentiment analysis techniques to the satisfaction recorded by expert physicians. In brief, two results appear to be of particular interest for the Crohn’s disease community of patients: (a) physicians tend to classify as neutral many posts that were previously classified as negative, and, (b) posts of patients that have been treated with Infliximab for a long time are never classified as negative by medical examiners. In addition, this short paper sets forth, with strength, a problem of methodological nature: is it possible to effectively analyze the big data of patient online posts without a stronger cooperation between data analysts and medical experts?

Efficient Adverse Drug Event Extraction Using Twitter Sentiment Analysis

Yang Peng, Melody Moh and Teng-Sheng Moh

Extensive clinical trials are required before a drug is placed on the market; yet it is difficult to discover all the side effects for any approved drugs. The United States Food and Drug Administration actively monitors approved medications to identify adverse events. The FDA Adverse Event Reporting System contains a database of adverse drug events (ADE) reported by the healthcare providers and consumers. The pervasive online social networks, such as Twitter, can provide additional information ADE. Concurrently, advancements in social media technology have resulted in the booming of massive public data; the availability of these huge datasets offers numerous research opportunities for extracting ADEs.

Towards this purpose, in this paper a simple, effective computation pipeline is proposed, which uses simple drug-related classification and sentiment analysis to extract ADEs on Twitter. The pipeline is described in detail, and is implemented into an automatic process. Experiments are carried out based on -months of Twitter data collected. Comparing with an existing pipeline, the new design is able to successfully capture times more valid ADEs, among them % are new ADEs. The proposed method may be applied to other areas such as food, beverages, and other daily consumer products for identifying side effects and user opinions.

HIBIB -- Session 2

A Hybrid Statistical and Semantic Model for Identification of Mental Health and Behavioral Disorders using Social Network Analysis

Madan Krishnamurthy, Khalid Mahmood and Pawel Marcinek

The advent of social networking and open health web forums such as PatientsLikeMe, WebMD, ehealth forum etc. have provided avenues for social user data that can prove instrumental in suggesting futuristic trends in healthcare. Homophily in social networks is a vital contributor for analyzing patterns for medical conditions, diagnosis and treatment options. Since, members with similar medical issues contribute to a common discussion pool; this offers a rich source of information that can be utilized. This paper intends to explore growing trends in Mental Health and Behavioral Studies (MHB) which lays emphasis on co-existing conditions resulting in comorbidity. We present a novel approach where personality traits inferred from unstructured text of patients and general social users are compared via statistical analysis. This is achieved by our Psychiatric Disorder Determination (PDD) algorithm. Further, Social media data of users showing personality traits of patients is subjected to semantic based text classification using Natural Language Processing (NLP) and Ontology Based Information Extraction (OBIE) in our Addiction Category Determination (ACD) algorithm. This provides categorization of user journals to common topics of discussion by referring to ontologies DBpedia, Freebase and YAGOs. The final category hence obtained can be predicted to be a trending subject of concern for users with Psychiatric disorders developing Addictive behavioral personalities.

A Personal Health Recommender System Incorporating Personal Health Records, Modular Ontologies, and Crowd-Sourced Data

Hengyi Hu, Adam Elkus and Larry Kerschberg

We present an architecture for a Personal Health Recommender System (PHRS) that begins with a person's personal electronic health record (PEHR) and augments it by combining crowd-sourced data mined for symptoms, diseases, treatments and best practices, together with authoritative sources and curated domain ontologies. This novel approach is patientcentric in that the PEHR contains the patient's health stats, specific symptoms, diagnoses, illnesses, medications and treatment plans. By accessing anonymized crowd-sourced data for similar cases, the patient and healthcare providers can ascertain the best course of treatment. In addition, the data in the PEHR can be classified according to authoritative ontologies using Semantic Web services. A modular ontology may be constructed for each of the patient's illnesses, and then they may be combined into an integrated ontology by patient or illness. As the PEHR is populated with more data, the ontology may evolve, guided by resources and services available on the Semantic Web.

A Smart Phone Application in Improving Healthy Lifestyles and Health Outcomes for School-age Children with Asthma

Chia-Tung Wu, Yu-Fen Tzeng, Te-Wei Ho, Shyh-Wei Chen, Bih-Shya Gau, Feipei Lai and Hung-Yu Chiu

Asthma is a complex multifactorial disease, increasing evidence shows that environmental and lifestyle factors may modify epigenetic mechanism of asthma development. Hence, the monitoring of asthma is essential for disease control and management. As mobile health technology has been widely applied in the prevention and management of health problem, smart phone applications (apps) providing social connection offer a potentially powerful approach to behavioral change through delivering convenient tailored intervention. However, there is limited evidence supporting apps can improve asthma self-management. Hence, the purposes of this study were to () develop an age-appropriate tailored KidsHealth app to improve healthy lifestyles for school-age children with asthma; () evaluate the feasibility and accessibility of interventions of the app; and () investigate the effectiveness of the app in improving healthy behaviors and outcomes for the users.

Mining Electronic Physical Records, A Trial

Xiaohui Wei, Wenyang Zou and Shang Gao

In China, the rapid development of medical informatization and the increasing public health awareness greatly stimulate the growth of Electronic Physical Records (EPR). This paper focuses on analyzing and mining electronic physical records to provide constructive suggestions for people's lifestyles using data of thirty thousand individual records in within the nine-month period from April to December, at China-Japan Union Hospital of Jilin University. During the analyzing process, numerous challenges were faced such as data storage, data noise and data incompletion etc. Thus we designed a framework as the resolution to these problems, and utilized data mining methods including association rules to find the characteristics of all ages in Changchun and gained some satisfactory results such as the abnormal BMI increasing with age. We also utilized the community discovery algorithm to explore the relationship between abnormal physical indicators. However, the result is not satisfactory and the module cant even reach .. To solve the problem of information isolation in China, we also implement a data platform.

HIBIB -- Session 3

Machine Learning Approach for Distinction of ADHD and OSA

Kuo-Chung Chu, Hsin-Jou Huang and Yu-Shu Huang

the purpose of this study is to find an efficient way to discriminate between Attention-deficit/ hyperactivity disorder (ADHD) and Obstructive sleep apnea (OSA). The study collected children (aged - years) data between and , who were divided into three groups, ADHD, OSA and a combination of ADHD and OSA. Each group based on the doctor's determination, using the DSM-IV diagnostic standards. The data included four questionnaires as follow: CBCL, DBRS, OSA- and CSHQ. In order to speed up the whole process of clinical diagnosis classification, we train and test three machine learning models to find the best way to help clinical doctor to diagnosis. The study results indicate that in all of subscale items, there were item show significantly difference among three subgroups, especially in the CBCL. Our results also show that CART model has better computational efficiency than CHAID and Neural Network model for subgroups classification.

A Fuzzy Model for Friendship Prediction in Healthcare Social Networks

Zeineb Dhouioui, Helmi Tlich, Radhia Toujeni and Jalel Akaichi

With the proliferation of social networks and their popularity especially Facebook and Twitter, healthcare related issues arise. Many users seek to follow and discuss similar disease experience. Thus,

it sounds promising to develop a friendship prediction system for healthcare purposes. Moreover, due to the dynamic aspect of social networks link (friendship) prediction has attracted the attention of many researchers. Link prediction consists on inferring probable links that may occur in the next time-stamp. Shortcomings of this task reside on that links are predicted according to only one time period and advanced privacy settings are behind the invisibility of interactions. In this work, we present an overview of some existing link prediction methods. We mainly propose a new approach to predict possible friendship between individuals and also we integrate fuzzy logic to deal with the lack of precision and the vagueness in the similarity between two individuals. Finally, in order to validate our solution, we use real data. The experiments show encouraging results. Comparing with crisp approaches, fuzzy method seems more effective and shows more accurate predictions.

DyNo Session 4

Aging Data in Dynamic Graphs: A Comparative Study

Anita Zakrzewska and David A. Bader.

Dynamic graphs are used to represent changing relational data. In order to create a dynamic graph representing relationships or interactions over time, it is necessary to choose a method of adding new data and removing, or otherwise deemphasizing, past data to decrease its influence. In particular, the question of aging edges is new to dynamic graphs and has not been thoroughly studied. In this work, we address the problem of aging vertices and edges to create a dynamic graph from a stream of temporal data. We provide two new methods, active vertex and active edge, and also evaluate two methods from the literature, sliding window and weight decay. By analyzing various properties of the dynamic graphs created by each aging method, we provide practitioners with quantitative comparisons. We find several interesting similarities and differences. The active vertex and weight decay methods reduce the variability over time of several vertex level measures compared to sliding window and active edge. This means that in practice, active vertex or weight decay may be more useful if graph stability is preferred, while sliding window or active edge may be preferred if the graph should be sensitive to changes in the underlying data stream. Each method also differently affects global measures. The most connected graph is produced by active vertex, while the most disconnected by weight decay. We observe that despite the differences, the graphs produced by each method experience similar types of changes at similar points in time.

Detecting Overlapping Community Hierarchies in Dynamic Graphs

Pascal Held and Rudolf Kruse

Community and cluster detection is a popular field of social network analysis. Most algorithms focus on static graphs or series of snapshots. In this paper we present an hierarchical algorithm, which detects communities in dynamic graphs. The method is based on the shortest paths to high-connected nodes, so called hubs. Due to local message passing, we can update the clustering results with low computational effort. The used hierarchy allows to process the community detection without setting any parameters before. After processing it provides different cluster levels based on the selected threshold. The presented algorithm is compared with the Louvain method on large-scale real-world datasets with given community structure. The detected community structure is compared to the given with NMI scores. The advantage of the algorithm is the good performance in dynamic scenarios.

Assessment of Effectiveness of Content Models for Approximating Twitter Social Connection Structures

Kuntal Dey, Sahil Agrawal, Rahul Malviya and Saroj Kaushik

This paper explores the social quality (goodness) of community structures formed across Twitter users, where social links within the structures are estimated based upon semantic properties of user-generated content (corpus). We examined the overlap of the community structures of the constructed graphs, and followership-based social communities, to find the social goodness of the links constructed. Unigram, bigram and LDA content models were empirically investigated for evaluation of effectiveness, as approximators of underlying social graphs, such that they maintain the community social property. Impact of content at varying granularities, for the purpose of predicting links while retaining the social community structures, was investigated. discussion topics, spanning over Twitter events, were used for experiments. The unigram language model performed the best, indicating strong similarity of word usage within deeply connected social communities. This observation agrees with the phenomenon of evolution of word usage behavior, that transform individuals belonging to the same community tending to choose the same words, made by [], and raises a question on the literature that use, without validation, LDA for content-based social link prediction over other content models. Also, semantically finer-grained content was observed to be more effective compared to coarser-grained content.

A Holistic Approach for Predicting Links in Coevolving Multiplex Networks

Alireza Hajibagheri, Gita Sukthankar and Kiran Lakkaraju

Networks extracted from social media platforms frequently include multiple types of links that dynamically change over time; these links can be used to represent dyadic interactions such as economic transactions, communications, and shared activities. Organizing this data into a dynamic multiplex network, where each layer is composed of a single edge type linking the same underlying vertices, can reveal interesting cross-layer interaction patterns. In coevolving networks, links in one layer result in an increased probability of other types of links forming between the same node pair. Hence we believe that a holistic approach in which all the layers are simultaneously considered can outperform a factored approach in which link prediction is performed separately in each layer. This paper introduces a comprehensive framework, MLP (Multiplex Link Prediction), in which link existence likelihoods for the target layer are learned from the other network layers. These likelihoods are used to reweight the output of a single layer link prediction method that uses rank aggregation to combine a set of topological metrics. Our experiments show that our reweighting procedure outperforms other methods for fusing information across network layers.

Modeling the Joint Dynamics of Relational Events and Individual States

Aaron Schecter and Noshir Contractor

Networks evolve at multiple levels; edges or relations may change, and the characteristics of the individuals within the network may change as well. Often these processes are intertwined, and in order to study them, statistical models must be developed that account for coevolution of multiple network components. The increased availability of continuous time network data has prompted new models for network inference such as the relational event framework. While network data may be continuously observable, the characteristics or state of an individual can still only be observed periodically. This type of panel data can be readily analyzed using actor-oriented models, but these methods do not accommodate continuous network data. We propose a model that integrates the relational event framework with actor-oriented models for behavioral change, allowing us to model the joint dynamics of relational events and individual states. This composite model preserves the advantages of each

method, while leveraging the richer information available in relational event data. We apply our model to datasets collected from virtual team experiments to highlight the utility of our method.

DyNo Session 1

The Haka Network: Evaluating Rugby Team Performance with Dynamic Graph Analysis

Paolo Cintia, Michele Coscia and Luca Pappalardo

Real world events are intrinsically dynamic and analytic techniques have to take into account this dynamism. This aspect is particularly important on complex network analysis when relations are channels for interaction events between actors. Sensing technologies open the possibility of doing so for sport networks, enabling the analysis of team performance in a standard environment and rules. Useful applications are directly related for improving playing quality, but can also shed light on all forms of team efforts that are relevant for work teams, large firms with coordination and collaboration issues and, as a consequence, economic development. In this paper, we consider dynamics over networks representing the interaction between rugby players during a match. We build a pass network and we introduce the concept of disruption network, building a multilayer structure. We perform both a global and a microlevel analysis on game sequences. When deploying our dynamic graph analysis framework on data from rugby matches, we discover that structural features that make networks resilient to disruptions are a good predictor of a team's performance, both at the global and at the local level. Using our features, we are able to predict the outcome of the match with a precision comparable to state of the art bookmaking.

Mining Social Interactions in Privacy-preserving Temporal Networks

Federico Musciotto, Saverio Delpriori, Paolo Castagno and Evangelos Pournaras

The opportunities to empirically study temporal networks nowadays are immense thanks to Internet of Things technologies along with ubiquitous and pervasive computing that allow a real-time finegrained collection of social network data. This empowers data analytics and data scientists to reason about complex temporal phenomena, such as disease spread, residential energy consumption, political conflicts etc., using systematic methologies from complex networks and graph spectra analysis. However, a misuse of these methods may result in privacyintrusive and discriminatory actions that may threaten citizens' autonomy and put their life under surveillance. This paper studies highly sparse temporal networks that model social interactions such as the physical proximity of participants in conferences. When citizens can self-determine the anonymized proximity data they wish to share via privacy-preserving platforms, temporal networks may turn out to be highly sparse and have low quality. This paper shows that even in this challenging scenario of privacy-by-design, significant information can be mined from temporal networks such as the correlation of events happening during a conference or stable groups interacting over time. The findings of this paper contribute to the introduction of privacy-preserving data analytics in temporal networks and their applications.

Narrative Smoothing: Dynamic Conversational Network for the Analysis of TV Series Plots

Xavier Bost, Vincent Labatut, Serigne Gueye and Georges Linares

Modern popular TV series often develop complex storylines spanning several seasons, but are usually watched in quite a discontinuous way. As a result, the viewer generally needs a comprehensive summary of the previous season plot before the new one starts. The generation of such summaries requires first to identify and characterize the dynamics of the series subplots. One way of doing so is to

study the underlying social network of interactions between the characters involved in the narrative. The standard tools used in the Social Networks Analysis field to extract such a network rely on an integration of time, either over the whole considered period, or as a sequence of several time-slices. However, they turn out to be inappropriate in the case of TV series, due to the fact the scenes showed onscreen alternatively focus on parallel storylines, and do not necessarily respect a traditional chronology. In this article, we introduce narrative smoothing, a novel, still exploratory, network extraction method. It smooths the relationship dynamics based on the plot properties, aiming at solving some of the limitations present in the standard approaches. In order to assess our method, we apply it to a new corpus of popular TV series, and compare it to both standard approaches. Our results are promising, showing narrative smoothing leads to more relevant observations when it comes to the characterization of the protagonists and their relationships. It could be used as a basis for further modeling the intertwined storylines constituting TV series plots.

MSNDS Session 1

Knowledge flow of biomedical informatics domain: position-based co-citation analysis approach

Kuo-Chung Chu and Chun-Cheng Yeh

A traditional co-citation analysis may ignore the structure of articles. When reference was cited in different level, such as, chapter level, the contribution will be different. In this study, we define the strength of chapter level, different than the traditional co-citation analysis; this study investigated the situation of the citation by chapter level. We use the cosine similarity to measure the similarity between content based co-citation, cluster analysis to verify results and visualization tool to present.

Deep Learning for Financial Sentiment Analysis on Finance News Providers

Min-Yuh Day and Chia-Chou Lee

Investors have always been interested in stock price forecasting. Since the development of electronic media, hundreds pieces of financial news are released on different media every day. Numerous studies have attempted to examine whether the stock price forecasting through text mining technology and machine learning could lead to abnormal returns. However, few of them involved the discussion on whether using different media could affect forecasting results. Financial sentiment analysis is an important research area of financial technology (FinTech). This research focuses on investigating the influence of using different financial resources to investment and how to improve the accuracy of forecasting through deep learning. The experimental result shows various financial resources have significantly different effects to investors and their investments, while the accuracy of news categorization could be improved through deep learning.

The Effect of Customer Perceived Value on Relationship Quality Between Illustrator and Fans to Recommendation on Facebook

Min-Yuh Day and Wei-Chun Chuang

In recent years, along with the prevalence of social networking sites, the illustrators of Wretch have accordingly transferred to new community platform. These illustrator's fan pages have become popular through viral marketing. Many companies have spotted enormous business opportunities and then worked with illustrators to boost sales by combining illustrators and commercial products. By utilizing a research model of relationship quality, online word-of-mouth, purchase intention and perceived value, the purpose of study is to explore whether fans would purchase the product endorsed by their favorite illustrators or other peripheral products. The finding show that relationship quality is positively related

to online word-of-mouth and had an indirect effect on purchase intention. Online word-ofmouth and purchase intention are positively correlated. In addition, perceived value is positively related to relationship quality, online word-of-mouth, and purchase intention. The main contribution of this research is in proposing a new research model and also discovering the importance of customer perceived value to the hedonic value of products.

Forming a Research Team of Experts in Expert-Skill Co-occurrence network of Research News

Juan Yang, Mengxin Li, Bin Wu and Chenyang Xu

The team formation problem is required to find a group of individuals that can match the skills required by a collaborative task. Large-scale and comprehensive scientific research tasks need skilled experts from various fields to form a research team and work for it. This paper constructs a dataset and proposes team formation algorithms to find out research teams, which provides decision support for the research projects. The size of existing datasets is relatively small and fields of experts in it are less diversified. This paper extracts information of experts and skills from research news to construct a co-occurrence network with heterogeneous network structure. Based on the dataset, this work designs approximate algorithms regarding skill as the priority to find near optimum teams with provable guarantees. On heterogeneous structure, the proposed algorithms directly search requested skills to form the subgraph of team, which achieve significant improvement in time efficiency. Experimental results suggest that our methods can form the high-quality research team, and have better efficiently compared to naive strategies and scale well with the size of the data.

A Comparative Study of Social Network Classifiers for Predicting Churn in the Telecommunication Industry

Maria Oskarsdottir, Cristian Bravo, Wouter Verbeke, Carlos Sarraute, Bart Baesens and Jan Vanthienen

Relational learning in networked data has been shown to be effective in a number of studies. Relational learners, composed of relational classifiers and collective inference meth- ods, enable the inference of nodes in a network given the existence and strength of links to other nodes. These methods have been adapted to predict customer churn in telecommunication companies showing that incorporating them may give more accurate predictions. In this research, the performance of a variety of relational learners is compared by applying them to a number of CDR datasets originating from the telecommunication industry, with the goal to rank them as a whole and investigate the effects of relational classifiers and collective inference methods separately. Our results show that collective inference methods do not improve the performance of relational classifiers and the best performing relational classifier is the network-only link-based classifier, which builds a logistic model using link-based measures for the nodes in the network.

MSNDS Session 2

A Novel Approach for m-representative Skyline Query

Heng-Shiou Sheu, Yi-Chung Chen and Don-Lin Yang

Skyline queries are currently the most notable type of multi-criteria search algorithm. A skyline query returns all of the data points in a given a dataset that are not dominated by other data points. However, this type of query is limited by the fact that the number of results cannot be controlled. In some cases, this can result in an excessive number of results, whereas other cases result in an insufficient number of results. In this study, we propose a scheme referred to as m-representative skyline queries to provide control over the number of results that are returned. We also developed a naive algorithm and a sorted

algorithm to provide additional control over the search process. Experiment results demonstrate the efficacy of the proposed approach.

Toward Understanding the Cliques of Opinion Spammers with Social Network Analysis.

Chih-Chien Wang, Min-Yuh Day and Yu-Ruei Lin

Consumer generated product reviews are considered as more persuasive than commercial advertising, and are now an important message source to make purchase decision. Nevertheless, firms may purposely hire spammers to create fake reviews to promote their products and to demote products of their competitors. To create the opinion majority, firms may hire a group of spammers rather than just one or few individual spammers to write fake reviewers. These spammers may act as a group to support other spammers to create a social consensus or majority of opinions. In the study, we attempt to adopt a real case to analyze the social network of spammers by K-core and Clique analysis. Our research results show that the social connection among spammers is stronger than that among non-spammers. Moreover, K-cores and cliques can be used as cues to identify spammers.

Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews

Kamil Topal and Gultekin Ozsoyoglu

Movie ratings and reviews at sites such as IMDb or Amazon are commonly used by moviegoers to decide which movie to watch or buy next. Currently, moviegoers base their decisions as to which movie to watch by looking at the ratings of movies as well as reading some of the reviews at IMDb or Amazon. This paper argues that there is a better way: reviewers movie scores and reviews can be analyzed with respect to their emotion content, aggregated and projected onto a movie, resulting in an emotion map for a movie. One can then make a decision on which movie to watch next by selecting those movies having emotion maps with certain emotion map patterns desirable for him/her. This paper is a first step towards the above-listed scenario.

Investor Classification and Sentiment Analysis

Arijit Chatterjee and William Perrizo

The paper discusses the bias of investors and the affect it has on the volatility of the stocks in the market. We also show how sentiment analysis can be run on the pulled tweets and why we chose the Microsoft Azure sentiment analyzer over the other commercial sentiment analyzer tools. Finally, we provide some future direction where we plan to take this research forward and conclude with some closing remarks.

SNAA - S1

Personal Networks and Perception of Care

Julian Fares and Kon Shing Kenneth Chung

During the past decade there has been a growing research interest in the effects of social support, characterized by social relationships and affiliation, on health. As health management is largely a social process, social networks have been theorized to impact health outcomes. We find however, an important gap in the literature. Little attention has been given to how the social structure and social position impact cancer care experience. The National Cancer Patient Experience survey is used to demonstrate how network data can be extracted for the purpose of studying the impact of social network properties on perception of care. The results show that there is a significant difference in network

properties (density, betweenness, degree, closeness, efficiency, constraint) of patients who were treated as a whole person or as a cancer symptom. We believe that social networks can help in improving the future service quality for cancer patients.

Evolutionary Algorithm for Seed Selection in Social Influence Process

Micha? Weskida and Rados?aw Michalski

Nowadays, in the world of limited attention, the techniques that maximize the spread of social influence are more than welcomed. Companies try to maximize their profits on sales by providing customers with free samples believing in the power of word-of-mouth marketing, governments and non-governmental organizations often want to introduce positive changes in the society by appropriately selecting individuals or election candidates want to spend least budget yet still win the election. In this work we propose the use of evolutionary algorithm as a mean for selecting seeds in social networks. By framing the problem as genetic algorithm challenge we show that it is possible to outperform well-known greedy algorithm in the problem of influence maximization for the linear threshold model in both: quality (up to % better) and efficiency (up to times faster). We implemented these two algorithms by using GPGPU approach showing that also the evolutionary algorithm can benefit from GPU acceleration making it efficient and scaling better than the greedy algorithm. As the experiments conducted by using three real world datasets reveal, the evolutionary approach proposed in this paper outperforms the greedy algorithm in terms of the outcome and it also scales much better than the greedy algorithm when the network size is increasing. The only drawback in the GPGPU approach so far is the maximum size of the network that can be processed - it is limited by the memory of the GPU card. We believe that by showing the superiority of the evolutionary approach over the greedy algorithm, we will motivate the scientific community to look for an idea to overcome this limitation of the GPU approach - we also suggest one of the possible paths to explore. Since the proposed approach is based only on topological features of the network, not on the attributes of nodes, the applications of it are broader than the ones that are dataset-specific.

Analysis of Link Formation, Persistence and Dissolution in NetSense Data

Ashwin Bahulkar, Boleslaw Szymanski, Omar Lizardo, Yuxiao Dong, Yang Yang and Nitesh Chawla

We study a unique behavioral network data set (based on periodic surveys and on electronic logs of dyadic contact via smartphones) collected at the University of Notre Dame. The participants are a sample of members of the entering class of freshmen in the fall of whose opinions on a wide variety of political and social issues and activities on campus were regularly recorded—at the beginning and end of each semester— for the first three years of their residence on campus. We create a communication activity network implied by call and text data, and a friendship network based on surveys. Both networks are limited to students participating in the NetSense surveys. We aim at finding student traits and activities on which agreements correlate well with formation and persistence of links while disagreements is highly correlated with non-existence or dissolution of links in the two social networks that we created. Using statistical analysis and machine learning, we observe several traits and activities displaying such correlations, thus being of potential use to predict social network evolution.

SNAA - S2

Improving the Robustness of the Smart Grid using a Multi-Objective Key Player Identification Approach

R. Chulaka Gunasekara, Kishan Mehrotra and Chilukuri Mohan

The smart grid interconnects a power grid (network) and a communication network, and enables bi-directional flow of electricity and information. To prevent the cascading failures which occur when the disruptions in one network cause disruptions in the other network, robustness should be enhanced by increasing the number of links (edges) between the power grid and the information flow network. Given a budget which constrains the number of new links that can be added to 'strengthen' the network, the best strategy to determine where to add those new links remains an open research problem. This paper presents a multi-objective approach to identify the best locations in the power network where new links can be added, to improve the overall robustness of the smart grid when constrained by resource limitations. Simulation results show that substantially greater robustness is obtained by using this approach, when compared to other link addition algorithms.

Towards Social Network Analytics for Understanding and Managing Enterprise Data Lakes

Ashley Farrugia, Rob Claxton and Simon Thompson

We have built a tool for inspecting and managing data lakes. The motivations for creating this tool are) schema discovery (determining links pertinent to solving a data analysis problem),) discovering high risk links in data schemas that give rise to Information Security problems and) discovering high value relationships enabling data asset curation. The tool works by extracting metadata from the Hive database on a shared-tenancy instance of Hadoop, which contained a multi-terabyte real-world data asset. We use this metadata to calculate a graph of the relationships between the entities based on column matching. This allows us to apply Social Network Analysis (SNA) techniques in order to discover meaningful properties of the accumulated data. For example to extract previously unknown relationships between data entities. The challenges and the agenda for future research are also provided.

Hybrid Structure!Vbased Link Prediction Model

Fei Gao and Katarzyna Musial

In network science several topology–based link prediction methods have been developed so far. The classic social network link prediction approach takes as an input a snapshot of a whole network. However, with human activities behind it, this social network keeps changing. In this paper, we consider link prediction problem as a time–series problem and propose a hybrid link prediction model that combines eight structure-based prediction methods and self-adapts the weights assigned to each included method. To test the model, we perform experiments on two real world networks with both sliding and growing window scenarios. The results show that our model outperforms other structure–based methods when both precision and recall of the prediction results are considered.

Community Evolution in Multiplex Layer Aggregation

Brian Crawford, Ralucca Gera, Ryan Miller and Bijesh Shrestha

This research studies community detection in multiplex dark networks. Our method seeks to intelligently select appropriate layers for aggregation to approximate communities in the whole network, while reducing the impact of over-modeling the network. Community evolution is explored as layers of different types of information are added to the partial picture of the network. We determine the set of dominant layers needed to produce similar community partitions to the established ground truth aggregate network. The identification of dominant layers enhances the selection of which layers to choose for aggregation purposes. This reduces redundancy and noise, and increases the optimization of the available data to produce the desired network partitions. We use normalized mutual index (NMI), purity, density, and modularity for methodology evaluation and comparison metrics.

SNAST Session 1

Automatic Tattoo Image Registration System

Xuan Xu, Michael Martin and Thirimachos Bourlai

Surveillance systems are very important for law enforcement and military applications. Capturing a biometric modality at a distance and under difficult conditions is a very challenging process. While face or gait can be used to identify an individual in such application, tattoos can also help in the identification process whenever available. Tattoos are considered a soft biometric and in some scenarios may be the only clue that can be used to verify the identity of a suspect or to rule out a suspect. One of the major challenges in tattoo recognition systems is image registration, i.e. the alignment of one tattoo image to a reference image. Accurate registration can greatly improve recognition accuracy. In this paper, we propose a twolevel automatic tattoo registration and correction system based on SIFT descriptors and the RANSAC algorithm with a homography model. By using image quality index techniques and a postprocessing step (where we refine our original registration results by an automated correction process where outliers are first identified and then re-processed), our system is able to demonstrate accurate registration results. We tested our registration system using two tattoo image databases. The first one is the NISTTatt- C database with subjects collected under uncontrolled condition, and the second one is the new WVU tattoo database (WVU-Tatt) with subjects, which is collected under controlled conditions. Experimental results show that, first, we obtained % registration accuracy in both databases. Then, the effect of our registration process on tattoo recognition performance was assessed when using both the NIST-Tatt-C database where the accuracy improved from .% (no registration) to % (with registration) and the WVU-Tatt database where the accuracy improved from .% (no registration) to .% (with registration).

Dynamic Prediction & Estimation of Intentional Failures in HPCs

Antwan Clark, L. M. Tellez, S. Besse and J. M. Absher

High performance computing systems are becoming the norm for daily use and their applications are being known within academia, industry, and government sectors. However, the resilience of these systems is in question for their complex internal structure makes them difficult to trouble shoot – making them vulnerable to intentional failures. Our work addresses this topic by employing dynamic prediction and estimation strategies of observed failures for individual nodal components within the exascale class network. Experiments using a simulated HPC environment convey the efficacy of our process where the results can be directly applied for improved detection and monitoring of cyber anomalies.

SNAST Session 2

Identification of Extremism on Twitter

Yifang Wei, Lisa Singh and Susan Martin

Identifying extremist-associated conversations on Twitter is an open problem. Extremist groups have been leveraging Twitter () to spread their message and () to gain recruits. In this paper, we investigate the problem of determining whether a particular Twitter user engages in extremist conversation. We explore different Twitter metrics as proxies for misbehavior, including the sentiment of the user's published tweets, the polarity of the user's ego-network, and user mentions. We compare different known classifiers using these different features on manually annotated tweets involving the ISIS

extremist group and find that combining all these features leads to the highest accuracy for detecting extremism on Twitter.

Stream Clustering of Tweets

Sophie Baillargeon, Simon Halle and Christian Gagne

This paper proposes an approach to cluster social media posts. It aims at taking full advantage of this recent source of newsworthy information and at facilitating the work of users who need to monitor public events in real-time. The emphasis is on developing a stream clustering algorithm able to process incoming tweets. A first implementation of the algorithm, focusing on the tweets' text, was tuned and tested on a dataset of manually annotated messages. Results show that the algorithm produces a partition of tweets similar to the manual partition obtained from humans. In future work, we plan to extend this algorithm with additional features and integrate the resulting analytical capabilities to a real-time social media monitoring platform called CrowdStack.

Profiling individuals based on email analysis and ego networks - A Visualization Technique

Andreas Xenaros, Panagiotis Karampelas and Ioanna Lekea

Nowadays, communication between people is mediated by technology and more specifically via Internet either by using email or social networking sites. Since any online activity generates an electronic trace, creating an automated tool to collect and analyze the communication between people can be valuable for extracting useful information about their behavioral characteristics. Combining these characteristics, with the ego network of each person within the network, additional analysis can be performed to detect the sphere of influence of the individual, identify other individuals in the network with similar behavioral characteristics and so on. In this paper, an automated tool has been designed and developed with the purpose to identify potential malevolent persons inside the organization. With this tool, we are able to analyze the communication between the employees of any given organization, build the behavioral profile of each individual, as well as their ego network.

Sentiment-enhanced Multidimensional Analysis of Online Social Networks: Perception of the Mediterranean Refugees Crisis

Mauro Coletto, Andrea Esuli, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego and Chiara Renso

We propose an analytical framework able to investigate discussions about polarized topics in online social networks from many different angles. The framework supports the analysis of social networks along several dimensions: time, space and sentiment. We show that the proposed analytical framework and the methodology can be used to mine knowledge about the perception of complex social phenomena. We selected the refugee crisis discussions over Twitter as a case study. This difficult and controversial topic is an increasingly important issue for the EU. The raw stream of tweets is enriched with space information (user and mentioned locations), and sentiment (positive vs. negative) w.r.t. refugees. Our study shows differences in positive and negative sentiment in EU countries, in particular in UK, and by matching events, locations and perception, it underlines opinion dynamics and common prejudices regarding the refugees.

On the Effectiveness of Statistical Hypothesis Testing in Infrared-based Face Recognition in Heterogeneous Environments

Neeru Narang and Thirimachos Bourlai

In this work, our objective is to study the impact of statistical hypothesis tests for the purpose of improving heterogeneous face recognition (FR). A series of tests are conducted to find the most suitable

type of statistical analysis test (parametric vs. non-parametric). To conduct the experiments, we used a multi-spectral face database (visible and Near-IR) collected under challenging conditions, i.e. at night time and at four different standoff distances, namely ; ; and meters. Next, the selected statistical analysis test is used to find the statistical significance of; (i) image restoration, (ii) fusion of scores. First, GaborWavelets, Histogram of gradients (HOG) and Local binary patterns (LBP) feature descriptors are empirically selected. Then the statistical analysis reveals which descriptors result in higher recognition performance. Finally, statistical hypothesis tests are performed to explore the impact of data stratification (grouping of gallery and probe sets) in terms of ethnicity, gender. A set of face identification studies are performed. Experimental results suggest that our proposed image restoration approach, fusion schemes and the usage of stratification result in a significantly better performance results than the baseline, e.g. the rankone score is improved from % to % when using image restoration, to % when using fusion of scores and to % (i.e. in the case of testing FR accuracy only on the female Asian class) when employing database stratification.

SI

Toward Understanding Spatial Dependence on Epidemic Thresholds in Networks

Zesheng Chen

Social influence in online social networks bears resemblance to epidemic spread in networks and has been studied through epidemiological models. The epidemic threshold is a fundamental metric used to evaluate epidemic spread in networks. Previous work has shown that the epidemic threshold of a network is exactly the inverse of the largest eigenvalue of its adjacency matrix. In this work, however, we indicate that such a threshold ignores spatial dependence among nodes and hence underestimates the actual epidemic threshold. Focusing on regular graphs, we analytically derive a more accurate epidemic threshold based on spatial Markov dependence. Our model shows that the epidemic threshold indeed depends on the spatial correlation coefficient between neighboring nodes and decreases with the death rate. Through both analysis and simulations, we show that our proposed epidemic threshold incorporates a certain spatial dependence and thus achieves a greater accuracy in characterizing the actual epidemic threshold in regular graphs. Moreover, we extend our study to irregular graphs by conjecturing a new epidemic threshold and show that such a threshold performs significantly better than previous work.

Observations on the role of influence in the difficulty of social network control

Dave Mckenney and Tony White

Previous work introducing the idea of distributionbased network control determined that some seemingly similar networks can have significantly different levels of controllability. This work investigates these differences in controllability in more detail and finds that one of the driving factors behind controllability may be the influence dynamics within the network. These results suggest that existing structural heuristics for control set selection that do not take into account influence, such as the FAR heuristic used here and in previous network control works, can produce substandard control sets that result in poor controller performance. The development of an algorithm for control set selection based on network influence measurements is identified as an important direction of future work.

Estimating Influence of Social Media Users from Sampled Social Networks

Kazuma Kimura and Sho Tsugawa

Several indices for estimating the influence of social media users have been proposed. Most such indices are obtained from the topological structure of a social network that represents relations among social media users. However, several errors are typically contained in such social network structures because of missing data, false data, or poor node/link sampling from the social network. In this paper, we investigate the effects of node sampling from a social network on the effectiveness of indices for estimating the influence of social media users. We compare the estimated influence of users, as obtained from a sampled social network, with their actual influence. Our experimental results show that using biased sampling methods, such as sample edge count, is a more effective approach than random sampling for estimating user influence, and that the use of random sampling to obtain the structure of a social network significantly affects the effectiveness of indices for estimating user influence, which may make indices useless.

Stability of Certainty and Opinion in Influence Networks

Ariel Webster, Bruce Kapron and Valerie King This paper introduces two models for influence in networks, and presents some upper and lower bounds for time needed to reach stability in these models. The first, called the Majority Model, is an expansion on the “Democrats and Republicans Model" that uses cascades to initialize the influence network rather than randomly assigning each node an initial opinion. By slightly modifying a network introduced by Frischknect, Keller, and Wattenhofer [] to fit the specifications of the Majority Model, we show that Frischknecht et al.’s lower bound for stability of ⌧(n) on the Democrats and Republicans Model also holds in the Majority Model. The second model, called the Certainty Model, is the same as the Majority Model but with the addition of a variable for a node’s certainty in its own opinion. Each node weights the opinions of its neighbors by their respective certainties and moves to the mass center of all of these opinions. For the Certainty Model we obtain two upper bounds related to time to stability. The first is a bound of O(d) for the time to reach stability once all nodes have gained an opinion, where d is the diameter of the graph. The second is a bound of O(n) on the time required for all nodes to gain an opinion.

Influential User Detection on Twitter: Analyzing Effect of Focus Rate

Zeynep Zengin Alp and Sule Gunduz Oguducu

Social media usage has increased marginally in the last decade and it is still continuing to grow. Companies, data scientists, and researchers are trying to infer meaningful information from this vast amount of data. One of the most important target applications is to find influential people in these networks. This information can serve many purposes such as; user or content recommendation, viral marketing, and user modeling. Social media is divided into subcategories like where one can share photos (i.e. Instagram, Flickr), video or music (i.e. Youtube, Last.fm), restaurant suggestions like Foursquare, or text like Twitter. Twitter is more of an idea and news sharing media than other types of social media and it has a huge amount of public profiles. These features of Twitter make it a more interesting and valuable media to research on. In this paper, we are addressing to identify topical authorities/ influential users in Twitter. We provide a novel representation of users’ topical interests called focus rate. We incorporate nodal features into network features and introduce a modified version of Pagerank algorithm which efficiently analyzes topical influence of users. Experimental results show that focus rate of users on specific topics increase their influence scores and lead to higher information diffusion. We use also distributed computing environment which enables to work with large data sets. We demonstrate our results on Turkish Twitter messages. For the best of our knowledge, this is the first influence analysis on Twitter that is conducted for Turkish language.

An Empirical Evaluation Of Social Influence Metrics

Nikhil Kumar, Ruocheng Guo, Ashkan Aleali and Paulo Shakarian

Predicting when an individual will adopt a new behavior is an important problem in application domains such as marketing and public health. This paper examines the performance of a wide variety of social network based measurements proposed in the literature - which have not been previously compared directly. We study the probability of an individual becoming influenced based on measurements derived from neighborhood (i.e. number of influencers, personal network exposure), structural diversity, locality, temporal measures, cascade measures, and metadata. We also examine the ability to predict influence based on choice of classifier and how the ratio of positive to negative samples in both training and testing affect prediction results - further enabling practical use of these concepts for social influence applications.

Doctoral Forum and Posters: Session 1: Social Network Analysis and Society

Is it truly a -star Movie? Restoring the Movie's Truthful Rating

Weiyue Huang and Yong Yu

Authenticity is the key for online review sites. Due to the significant development of review sites, the reviews are now highly important to users, producers and other stakeholders. Driven by interest, some imposters begin to post fake reviews to promote or discredit target products. The fake reviews not only mislead the users but also damage the service provider’s credit. Current works mostly aim at classifying whether a specific review is fake or not, using context-based or user-based approaches. However, the aggregated rating of the product is viewer’s most concern. Therefore, we propose a novel task to restore the truthful rating and further tackle it by statistical and deep learning techniques. We also assemble and publish a movie-review dataset for this task.

Emerging Threats Abusing Phone Numbers Exploiting Cross-Platform Features

Srishti Gupta

Phone number, a unique identifier has emerged as an important Personally Identifiable Information (PII) in the last few years. Other PII like e-mail and online identity have been exploited in the past to launch phishing and spam attacks against them. The reach and security of a phone number provide a genuine advantage over e-mail or online identity, making it the most vulnerable attack vector. In this work, we explore the emerging threats that abuse phone numbers by exploiting crossplatform features. Given that phone number space hasn’t been extensively studied in the past, there is a dire need to understand the threat landscape and develop solutions to prevent its abuse.

GeoContext: Discovering Geographical Topics from Social Media

Elizabeth Williams

Social media is often useful for discovering contextual information that is difficult to find on traditional query-based search engines such as Google. For example, temporal events such as traffic incidents are often posted on social media due to the wide-reaching and real-time nature of social media platforms. Social media can also be used to model the sentiments and opinions of different geographical regions, as well as provide a platform for organizing social movements. In this paper, we give an overview of our system, GeoContext, which models a stream from Twitter into topics and analyzes the geographical locations of the topics. GeoContext includes methods for filtering a social media stream by keywords and location coordinates in order to provide more specific topics. GeoContext includes a geolocation

module, called GeoContext Locator, for predicting the locations of tweets that are not associated with explicit coordinates, in order to model topics in different locations.

Doctoral Forum and Posters: Session 2: Fundamentals of Social Network Analysis

Detecting and Mitigating Bias in Social Media

Fred Morstatter

Social media is an important data source. Every day, billions of posts, likes, and connections are created by people around the globe. By monitoring it we can observe important topics, as well as find new topics of discussion as they emerge. However, within this source of information there are natural forms of bias. Different aspects of the sites lend themselves to bias, such as varying features that restrict users. Additionally, the users themselves can be biased, such as the age-bias found in Twitter users. Finally, the way sites divulge their data can cause bias to those studying information produced on that site. In this forum we will discuss the different types of bias that can occur on social media data as well as different strategies to mitigate that bias.

Noise Removal and Structured Data Detection to Improve Search for Personality Features

Muhammad Fahim Uddin

This paper discusses part one of the main work in field of data science, mining and analytics. Family of algorithms is developed to predict the educational relevance of individuals' talents through lens of personality features (unstructured and semi-structured) and academic/career data. The big data (unstructured and semi-structured) contains lots of valuable information that can be mined and analyzed. However, such processing and utilization of data introduce challenges of dealing with noise (irrelevant, unnecessary and redundant data). Regardless of the nature of data processing and utilization for a given problem or an application, noise adds unnecessary time and cost. This paper briefly discusses the overall research work and then presents Noise Removal and Structured Data Detection (NR-and-SDD) algorithm and related math construct. NR-and- SDD detects the noise to reduce the processing cost and improve structured data detection in relevance of personality features. The given results show improved reliability and efficiency of NR and SDD processes. Related study is provided and paper is concluded with final remarks and future works.

Towards Using Subpattern Distributions in Social Network Analysis

Benjamin Cabrera

Determining the frequencies and the distribution of small subgraph patterns in a large input graph is an important part of many graph based mining tasks such as Frequent Subgraph Mining (FSM) and Motif Detection. Due to the exponential number of such graph patterns the interpretation of the mining results is mostly limited to finding unexpectedly frequent patterns, and in general identifying few particularly interesting patterns to then understand their function in the network. However, the full distribution of patterns itself encodes much more information about the underlying graph. Looking at this pattern distribution could be of particular interest in Social Network Analysis because social networks seem to have more random noise and less hard structural constraints compared to other types of graphs. This makes it unlikely to find meaning in only the most frequent patterns. In this paper we contribute in two ways to the usage of pattern distributions in networks analysis. First, we introduce two different sampling-based algorithms for computing the mentioned pattern distributions. Second we sketch two

original ideas which can be used to harness the information in pattern distributions for gaining insights on global and local properties of Social Networks.

Doctoral Forum and Posters:

Arguments and Interpretation in Big Social Data Analysis: A Survey of the ASONAM Community

Candice Lanius

Big social data is becoming an important part of human decision making around the world. With the high stakes of decisions based on technical systems, it is important to evaluate the role of researchers in shaping that shared future. I present the results of a survey of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining participants who performed big social data analyses. The goal was to understand how data scientists use interpretation to complete their projects and how they communicate results to their audience. By looking at research design as both a technical roadmap and an argument, results generated from social media data sets can be evaluated for their quality. Ultimately, these results will assist in the creation of field-dependent evaluation standards than can be used by big social data researchers.

Identifying the Impact of Friends on their Peers Academic Performance

Philip Scanlon

Historically data collection in the research process involves either surveys, interviews or observation, or any combination of all three. Recent developments in the area of formative educational methods have enabled other data collection options. Data sources now available include logs from University Virtual Learning Environments (VLEs), E-learning and many other knowledge management systems. Datasets harvested from these sources are less susceptible to the inherent biases introduced through the intervention of human interpretation. Data is often structured, complete and traceable. The research in this paper aims to utilise one of these unique digital datasets which represents the footprints created by student activities within a university environment and through Social Network Analysis to identify their influences within peer groups.

A Platform for Identifying Experts and Paper Retrieval in Citation Networks

Qianwei Wang, Qianshan Yu and Xinyue Yu

Efficient organization and analysis of academic information has many advantages. Most scholar retrieval systems appeared these years can perform keyword-based paper search. However, performing large-scale expert and paper retrieval is an intractable problem. Here we present a platform that can not only reduce the workload of researchers when searching academic literature, but also promote academic communication. In this paper, we introduced a novel community partition algorithm specific to deal with large-scale citation network, the Large-scale Citation Network Partition Algorithm (LCNPA). We demonstrate the construction of the platform, and illustrate the implemented functions and instructions.

Gamification for Informal Terms Lexicon Building

Fernando Henrique Calderon Alvarado and Yi-Shin Chen

Image tagging approaches have gained popularity in recent years. Advances in social computing research have enabled a faster completion of this task which before was known to be tedious. Most of

which focus on the image-label relationship, leading to a vast amount of image repositories with corresponding text descriptors. There are several other collections such as dictionaries, lexicons or ontologies that aid different tasks the domains of text mining, Natural Language Processing and the different applications that come with it. Many of the research in this area is oriented towards social network analysis. These collections are based on formal ways of expression and unfortunately social networks content is not always formal. There is an increasing use of informal, many times regional language popularly called slang. We propose a Game With a Purpose (GWAP) system to facilitate the collection of informal terms. Leveraging on social interactions we seek to obtain a lexicon that is more suitable for social media related analysis.

Social Network of Software Development at GitHub

William Leibzon

This paper looks at organization of software development teams and project communities at GitHub. Using social network analysis several open-source projects are analyzed and social networks of users with ties to a project are shown to have some scale-free properties. We further show how to find core development group and a network metric is introduced to measure collaboration among core members, corresponding to if a project is healthy and more likely to be successful.

Posters/Demos Madness Session

BullyBlocker: Towards the Identification of Cyberbullying in Social Networking Sites

Yasin Silva, Christopher Rich and Deborah Hall

Cyberbullying is the deliberate use of online digital media to communicate false, embarrassing, or hostile information about another person. It is the most common online risk for adolescents and well over half of young people do not tell their parents when it occurs. While there have been many studies about the nature and prevalence of cyberbullying, there has been relatively less work in the area of automated identification of cyberbullying in social media sites. The focus of our work is to develop an automated model to identify and measure the degree of cyberbullying in social networking sites, and a Facebook app for parents, built on this model, that notifies them when cyberbullying occurs. This paper describes the challenges associated with building a computer model for cyberbullying identification, presents key results from psychology research that can be used in such a model, describes an initial model and mobile app design for cyberbullying identification, and describes key areas of future work to improve upon the initial model.

Learning Triadic Influence in Large Social Networks

Chenhui Zhang, Sida Gao, Jie Tang, Tracy Xiao Liuy, Zhanpeng Fang and Xu Chen

Social influence has been a widely accepted phenomenon in social networks for decades. In this paper, we study influence from the perspective of structure, and focus on the simplest group structure—triad. We analyze two different genres of behavior: Retweeting on Weibo and Paying on CrossFire. We have several intriguing observations from these two networks. First, different internal structures of one's friends exhibit significant heterogeneity in influence patterns. Second, the strength of social relationship plays an important role in influencing one's behavior, and more interestingly, it is not necessarily positively correlated with the strength of social influence. We incorporate the triadic influence patterns into a predictive model to predict user's behavior. Experiment results show that our method can significantly improved the prediction accuracy.

Local Community Detection in Multilayer Networks

Roberto Interdonato, Andrea Tagarelli, Dino Ienco, Arnaud Sallaberry and Pascal Poncelet

The problem of local community detection refers to the identification of a community starting from a query node and using limited information about the network structure. Existing methods for solving this problem however are not designed to deal with multilayer network models, which are becoming pervasive in many fields of science. In this work, we present the first method for local community detection in multilayer networks. Our method exploits both internal and external connectivity of the nodes in the community being constructed for a given seed, while accounting for different layer-specific topological information. Evaluation of the proposed method has been conducted on realworld multilayer networks.

Node-Centric Detection of Overlapping Communities in Social Networks

Yehonatan Cohen, Danny Hendler and Amir Rubin

We present NECTAR, a community detection algorithm that generalizes Louvain method's local search heuristic for overlapping community structures. NECTAR chooses dynamically which objective function to optimize based on the network on which it is invoked. Our experimental evaluation on both synthetic benchmark graphs and real-world networks, based on ground-truth communities, shows that NECTAR provides excellent results as compared with state of the art community detection algorithms.

Social Network Change Detection Using a Genetic Algorithm Based Back Propagation Neural Network Model

Ze Li, Duo-Yong Sun, Jie Li and Zhan-Feng Li

Changes in social networks may reflect an under-lying significant events or behaviors within an organization. Detecting these changes effectively and efficiently could have the potential to enable the early warning, and faster response to both positive and negative organizational activities. In this paper, we use a genetic algorithm based back propagation (GABP) neural network model to quantitatively determine if and when a change has occurred. By selecting network measures as input and dynamic network behavior types as output, we get the GABP neural network model well trained. Then, this approach is applied to Enron social networks. The results indicate that this approach achieves higher detection precision.

Towards Predicting Academic Impact from Mainstream News and Weblogs: A Heterogenous Graph Based Approach

Mohan Timilsina, Brian Davis, Mike Taylor and Conor Hayes

The realization that scholarly publications are discussed and have influence on discourse outside scientific and academic domains has given rise to area of scientometrics called alternative metrics or ”altmetrics”. Furthermore, researchers in this field tend to focus primarily on measuring scientific activity on social media platforms such as Twitter, however these countbased metrics are vulnerable to gaming because they tend to lack concrete justification or reference to the primary source. In this collaboration with Elsevier, we extend the conventional citation graph to a heterogeneous graph of publications, scientists, venues, organizations and more authoritative media sources such as mainstream news and weblogs. Our approach consists of two parts: one is integrating the bibliometric data with the social data such as blogs, mainstream news. The other involves understanding how standard graph-based metrics can be used to predict the academic impact. Our result showed the computed graph-based metrics can reasonably predict the academic impact of early stage researchers.

Towards Sentiment Analysis for Mobile Devices

Johnnatan Messias, Joao P. Diniz, Elias Soares, Miller Ferreira, Matheus Araujo, Lucas Bastos, Manoel Miranda and Fabricio Benevenuto

The increasing use of smartphones to access social media platforms opens a new wave of applications that explore sentiment analysis in the mobile environment. However, there are various existing sentiment analysis methods and it is unclear which of them are deployable in the mobile environment. This paper provides the first of a kind study in which we compare the performance of sentence-level sentiment analysis methods in the mobile environment. To do that, we adapted these sentencelevel methods to run on Android OS and then we measure their performance in terms of memory usage, CPU usage, and battery consumption. Our findings unveil sentence-level methods that require almost no adaptations and run relatively fast as well as methods that could not be deployed due to excessive use of memory. We hope our effort provides a guide to developers and researchers interested in exploring sentiment analysis as part of a mobile application and can help new applications to be executed without the dependency of a server-side API.

Trivia Quiz Mining using Probabilistic Knowledge

Taesung Lee, Seung-Won Hwang and Zhongyuan Wang

Recent work suggests that providing unexpected information is an important factor for drawing user traffic. Such examples can be easily found in the “Did you know” section of the Wikipedia main page, the ESPN quiz, the Google Doodles, and the Bing main page. Inspired by these applications, we propose a novel trivia quiz mining asking unexpected questions for a given entity. We solve this problem by linking different types of social media as input and output, and mine unexpected properties based on prototype theory to mediate the input and the output media.

Troll Vulnerability in Online Social Networks

Paraskevas Tsantarliotis, Evaggelia Pitoura and Panayiotis Tsaparas

Trolling describes a range of antisocial online behaviors that aim at disrupting the normal operation of online social networks and media. Combating trolling is an important problem in the online world. Existing approaches rely on human-based or automatic mechanisms for identifying trolls and troll posts. In this paper we take a novel approach to the trolling problem: our goal is to identify the targets of the trolls, so as to prevent trolling before it happens. We thus define the troll vulnerability prediction problem, where given a post we aim at predicting whether it is vulnerable to trolling. Towards this end, we define a novel troll vulnerability metric of how likely a post is to be attacked by trolls, and we construct models for predicting troll-vulnerable posts, using features from the content and the history of the post. Our experiments with real data from Reddit demonstrate that our approach is successful in recalling a large fraction of the troll-vulnerable posts.

ARC: A Pipeline Approach Enabling Large-Scale Graph Visualization

Michael Ferron, Ken Q. Pu and Jaroslaw Szlichta

When working with a high volume relational database, is it possible to effectively provide a compact visualization of the tuples in that database? Data visualization techniques very often scale poorly with input volume, hindering attempts at providing a responsive, full picture of the relationships within data. We introduce a method of efficiently visualizing millions of tuples in a two-dimensional constrained space, providing a method for data to be visually analyzed at the tuple level.We achieve this by

applying a physics simulation on an embedded network, positioning tuples according to their representative node.

BullyBlocker: An App to Identify Cyberbullying in Facebook

Yasin N. Silva, Christopher Rich, Jaime Chon and Lisa M. Tsosie

Cyberbullying is the most common online risk for adolescents, and it has been reported that over half of young people do not tell their parents when it occurs. Cyberbullying involves the deliberate use of online digital media to communicate false or embarrassing information about another person. While previous work has extensively analyzed the nature and prevalence of cyberbullying, there has been significantly less work in the area of automated identification of cyberbullying, particularly in social networking sites. The focus of our work is to develop a computational model to identify and measure the intensity of cyberbullying in social networking sites. In this paper, we present and demonstrate BullyBlocker, an app that identifies instances of cyberbullying in Facebook and notifies parents when it occurs. This paper presents the most relevant characteristics of our initial cyberbullying identification model, key app design and implementation details, the demonstration scenarios, and several areas of future work to improve upon the initial model.

CredFinder: a Real-time Tweets Credibility Assessing System

Majed AlRubaian, Muhammad Al-Qurishi, Mabrook Al-Rakhami, Mohammad Mehedi Hassan and Atif Alamri

Lately, Twitter has grown to be one of the most favored ways of disseminating information to people around the globe. However, the main challenge faced by the users is how to assess the credibility of information posted through this social network in real time. In this paper, we present a real-time content credibility assessment system named CredFinder, which is capable of measuring the trustworthiness of information through user analysis and content analysis. The proposed system is capable of providing a credibility score for each user's tweets. Hence, it provides users with the opportunity to judge the credibility of information faster. CredFinder consists of two parts: a frontend in the form of an extension to the Chrome browser that collects tweets in real time from a Twitter search or a user-timeline page and a backend that analyzes the collected tweets and assesses their credibility.

Finding Requests in Social Media for Disaster Relief

Tahora H. Nazer, Fred Morstatter, Harsh Dani and Huan Liu

Natural disasters create an uncertain environment in which first responders face the challenge of locating affected people and dispatching aids and resources in a timely manner. In recent years, crowdsourcing systems have been developed to exploit the power of volunteers to facilitate humanitarian logistic efforts. Most of the current systems require volunteers to directly provide input to them and do not have the capability to benefit the large number of disaster-related posts that are published on social media. Hence, many social media posts in the aftermath of disasters remain hidden. Among these hidden posts are those that need immediate attention, such as requests for help. Hence, we have implemented a system that detects requests on Twitter using content and context of tweets.

GiveMeExample: Learning Confusing Words by Example Sentences

Chieh-Yang Huang and Lun-Wei Ku

The rapid growth of web source has changed language learning behavior. More and more people utilized web sources instead of paper books. However, the problem now is that it is overwhelming to

find useful information. In addition, when considering using different words, good example sentences demonstrating nuance among words are extremely helpful but learners can hardly find them as most web dictionaries contain explanations and examples for only a single word. To solve the problem, we proposed a system called GiveMeExample which can automatically search for the best example sentences of a group of confusing words. The proposed system learns the word usage model for each word in the confusing word group and a universal difficulty model for all sentences, in order to propose simple but clear example sentences for learners. Experiments show that the proposed approach can really provide the most useful example sentences for understanding confusing words. GiveMeExample is available at http://givemeexample. com/GiveMeExample.

MIDAS: Mental Illness Detection and Analysis via Social Media

Elvis Saravia, Chun-Hao Chang, Renaud Jollet De Lorenzo and Yi-Shin Chen

Mental illnesses rank as some of the most disabling conditions, affecting millions of people, across the globe. In general, the main challenge of mental disorders is that they remain difficult to detect on suffering patients. In an online environment, the challenge extends to the collection of patients data and the implementation of proper algorithms to assist in the detection of such illnesses. In this paper, we propose a novel data collection mechanism and build predictive models that leverage language and behavioral patterns, used particularly on Twitter, to determine whether a user is suffering from a mental disorder. After training the predictive models, they are further pre-trained to serve as the back-end for our demonstration, MIDAS. MIDAS offers an analytics web-service to explore several characteristics pertaining to user's linguistic and behavioral patterns on social media, with respect to mental illnesses.

Polinode: A Web Application for the Collection and Analysis of Network Data

Andrew Pitts

We introduce Polinode, an online tool for performing network analysis. Polinode is aimed at commercial and non-commercial users alike and supports both researchrelated use cases as well as teaching network analysis to students. One of its primary advantages is that it is web-based. It therefore doesn't require any software downloads and opens up new avenues for collaboration and the incorporation of online content into network analysis. With Polinode you can upload arbitrary network data for online visualisation and analysis and you can also use the built-in relationship-based survey functionality to collect network data from respondents.

POSN: A Privacy Preserving Decentralized Social Network App for Mobile Devices

Eric Klukovich, Esra Erdin and Mehmet Hadi Gunes

Social networking has influenced billions of users to interact and share information online with friends and family. As online interactions have become the norm and online platforms amassed user data, privacy concerns for the user data has increased. Decentralized architectures can provide better privacy to the users by removing the central authority but have performance issues in dissemination of the content. In this study, we present the Personal Online Social Network (POSN) app that implements a cloud-backed peer-to-peer decentralized OSN using mobile devices. In POSN, each user utilizes a storage cloud to store and distribute encrypted content to his/her friends. Direct key management allows the user to have fine-grained access control of the shared content, and protects the data from being accessed by third parties. The POSN app is available at https://github.com/posn/POSN-app.

PRO-Fit: Exercise with friends

Saumil Dharia, Vijesh Jain, Jvalant Patel, Jainikkumar Vora, Rizen Yamauchi, Magdalini Eirinaki and Iraklis Varlamis

The advancements in wearable technology, where embedded accelerometers, gyroscopes and other sensors enable the users to actively monitor their activity have made it easier for individuals to pursue a healthy lifestyle. However, most of the existing applications expect continuous commitment from the end users, who need to proactively interact with the application in order to connect with friends and attain their goals. These applications fail to engage and motivate users who have busy schedules, or are not as committed and self-motivated. In this work, we present PRO-Fit, a personalized fitness assistant application that employs machine learning and recommendation algorithms in order to smartly track and identify user's activity, synchronizes with the user's calendar, recommends personalized workout sessions based on the user's preferences, fitness goals, and availability. Moreover, PRO-Fit integrates with the user's social network and recommends "fitness buddies" with similar preferences and availability.

VirtualIdentity: Privacy Preserving User Profiling

Sisi Wang, Wing-Sea Poon, Golnoosh Farnadi, Caleb Horst, Kebra Thompson, Michael Nickels, Anderson Nascimento and Martine De Cock

User profiling from user generated content (UGC) is a common practice that supports the business models of many social media companies. Existing systems require that the UGC is fully exposed to the module that constructs the user profiles. In this paper we show that it is possible to build user profiles without ever accessing the user's original data, and without exposing the trained machine learning models for user profiling – which are the intellectual property of the company – to the users of the social media site. We present VirtualIdentity, an application that uses secure multi-party cryptographic protocols to detect the age, gender and personality traits of users by classifying their usergenerated text and personal pictures with trained support vector machine models in a privacy preserving manner.

A Centrality-based Measure of User Privacy in Online Social Networks

Ruggero G. Pensa and Gianpiero Di Blasi

The risks due to a global and unaware diffusion of our personal data cannot be overlooked when more than two billion people are estimated to be registered in at least one of the most popular online social networks. As a consequence, privacy has become a primary concern among social network analysts and Web/data scientists. Some studies propose to "measure" users' profile privacy according to their privacy settings but do not consider the topological properties of the social network adequately. In this paper, we address this limitation and define a centrality-based privacy score to measure the objective user privacy risk according to the network properties. We analyze the effectiveness of our measures on a large network of real Facebook users.