University of Michigan
Computer Science and Engineering
and School of Information

Workshop on Data, Text, Web, and Social Network Mining

Friday, April 23, 2010

Schedule

9:30 a.m. **Kickoff, 1670 CSE**

9:30 to 9:40 a.m. **Intro and Welcome, 1670 CSE**

Dragomir Radev, Associate Professor, CSE, Professor, SI
H.V. Jagadish, Professor, CSE
Farnam Jahanian, Professor and Chair, CSE
Raghu Ramakrishnan, Yahoo!

9:40 to 11:00 a.m. **SESSION I – Lab Overviews (1), 1670 CSE**

Eytan Adar, Assistant Professor, SI and CSE

The Temporal-Informatics research group (*http://www.cond.org*) seeks to understand and enhance interactions with temporally varying, Web-scale information. "Understanding" involves the study, and modeling, of the evolution of information both in terms of the content itself and the production and consumption behaviors of large populations. To enhance interactions we implement solutions that provide better ways to access and manipulate dynamic information. Research methodologies range from text and log mining, information retrieval, human-computer interaction, network analysis and visualization (with some side trips into security and peer-to-peer systems).

H.V. Jagadish, Professor, CSE

The group's research focuses on managing the complexity of data in contexts such as the web and biomedicine. The goal of the lab is to improve the usability of databases for the end user, specifically in the areas of data management, querying and analysis. Interests in data management include schema evolution, provenance management and data integration challenges for biological databases. Efforts in database querying involve the creation of usable query models and interfaces for structured information. Research in supporting effective data analysis addresses summarization of database schemas, enabling information discovery and visualization of structured and link-based data.

Kristen LeFevre, Assistant Professor, CSE

Our group works primarily at the intersection of database systems, data mining, and security/privacy. In this talk, I will give a short overview of two current project areas: data use auditing (understanding a posteriori how sensitive private data has been used, and responding appropriately), and privacy for social networks.

Fan Meng, Assistant Research Professor, Psychiatry Department and Molecular and Behavioral Neuroscience Institute

Our research is in the area of high throughput biomedical data analysis (http://brainarray.mbni.med.umich.edu). We are interested in developing more efficient solutions to facilitate hypothesis development for biomedical researchers. Two of our focuses are 1) the development of better entity recognition methods for biomedical literature 2) solutions for more effective exploration of Medline literature together with experimental data.

Chris Miller, Assistant Professor, Astronomy

The goals of the INternational Computational Astrostatistics group (INCA, http://incagroup.org) are to develop and apply novel statistical methods to inference problems in astronomy and cosmology, with an emphasis on computational non-parametric approaches. The size and complexities of astronomical data often present statistical as well as computational challenges. Recent efforts include the development of non-parametric techniques to quantify the features of the data and models which describe the evolution and content of the Universe. We also study the diversity of the shapes and colors of the galaxy populations in order to explain how galaxies how evolved over time. Our techniques are being developed in the context of the next generation large-scale astronomical datasets (e.g., the Large Synoptic Survey Telescope (http://www.lsst.org).

XuanLong Nguyen, Assistant Professor, Statistics

Long Nguyen and his research group are interested in developing computational and statistical models, machine learning and optimization methods for structured data and distributed systems. Examples of on-going projects include distributed statistical inference and variational inference methods; nonparametric Bayesian modeling for functional and spatio-temporal data; image, signal and text processing; Detection, tracking and estimation applications in distributed systems and sensor networks; Applications of modeling and computation in the environmental science.

Maggie Levenstein, Associate Research Scientist, Institute for Social Research

The Michigan Census Research Data Center (MCRDC) provides qualified researchers with approved projects with access to restricted U.S. Census Bureau, National Center for Health Statistics, and Agency for Healthcare Research and Quality business and household microdata. Research topics include: Determinants of productivity growth, Business entry, exit, and employment decisions, Impact and determinants of international trade and outsourcing, Migration, Aging of the labor force, Health disparities, Impact of vouchers on educational outcomes, Disability and labor supply, Synthetic data and local area estimates. MCRDC is a unit of the Institute for Social Research. Its executive director is Maggie Levenstein and its Census Administrator is J. Clint Carter. URL: http://www.isr.umich.edu/src/mcrdc/

Qiaozhu Mei, Assistant Professor, SI and CSE

Qiaozhu Mei's research group focuses on the theory and applications of information retrieval and data mining. Specific research topics include probabilistic models and scalable inference algorithms for text, network, and user behavior data, language modeling in retrieval, contextual text mining, statistical topic models, graph-based regularization and smoothing, and personalization and diversity in ranking. Specific applications include intelligent web search, mining and summarizing social media and social networks, sentiment and opinion analysis, knowledge discovery from scientific literature, and mining electronic health information and records.

11:00 to 11:10 a.m. **Break**

11:10 to Noon **SESSION II – LAB Overviews (2), 1670 CSE**

Michael Cafarella, Assistant Professor, CSE

Professor Cafarella's research addresses different aspects of Web data management, including how to extract structured data from unstructured documents, and how to perform such work efficiently. Recent projects include an effort to extract an enormous database collection from crawled HTML tables, a system for quickly integrating data from multiple online sources, and a relational-style optimizer for MapReduce programs.

Dragomir Radev, Associate Professor, CSE, Professor, SI

Dragomir Radev’s research group, CLAIR (http://clair.si.umich.edu), is a place to study computational linguistics and information retrieval from a statistical and networks perspective. Some recent and current projects include the study of speaker centrality in political speeches, automatic summarization of news and scientific articles, open-domain question answering, language networks and lexical centrality, information extraction for the life sciences, the evaluation of machine translation, information obsolescence on the web, computational sociolinguistics, time-dependent topic modeling, citation analysis, semi-supervised text classification, graph-based methods for natural language processing, identifying gene-disease associations, etc.
Lab members: Ahmed Hassan (PhD student), Arzucan Ozgur (PhD student), Pradeep Muthukrishnan (PhD student), Vahed Qazvinian (PhD student), Amjad abu Jbara (PhD student), Yang Liu (MS student).

Gus Rosania, Associate Professor, Pharmaceutical Sciences

The Rosania research group is interested in developing computational tools that address several unmet research needs in pharmaceutical sciences. These computational tools include 1) multiscale, predictive models of drug transport in living organisms; 2) machine vision for quantifying the distribution of bioimaging probes in living cells; 3) cheminformatics tools for optimizing the targeting mechanism of drug-like molecules; 4) image-based data mining tools for mining the chemical literature; 5) visualization of large image datasets. We are primarily involved in mechanistic, hypothesis-driven, pharmaceutically-relevant projects at the interface of chemistry, biology, computer science. The interdisciplinary nature of these projects has evolved into several different collaborations with computational researchers in various departments within the University of Michigan, including bioinformatics, statistics and engineering.

Lada Adamic, Assistant Professor, SI and Complex Systems

The NetSI group (http://netsi.org) studies information dynamics in networks: how information diffuses, how it can be found, and how it influences the evolution of a network's structure. Recent projects include: identifying and motivating expertise in online question-answer forums, predicting information diffusion over social networks in virtual worlds, optimizing resource allocation over networks to prevent diffusion spread, and quantifying the effect of anonymity and reciprocity on truthfulness of online trust ratings.
Lab members: David Huffaker (postdoc), Jiang Yang (PhD student), Eytan Bakshy (PhD student), Matthew Simmons (PhD student), Xiao Wei (MSI student), Chun Yuen Teng (MSI student).

Yilu Murphey, Professor, ECE Department, UM-Dearborn

One of the major research activities at the Intelligent Systems Lab (ISL) at the ECE Department, UM-Dearborn, is to apply machine learning techniques to text data mining. We focus on unstructured text documents, namely documents do not follow any standard format and grammar. Examples of such documents are casual descriptions such as customer descriptions of vehicle problems, emails and web blogs. This type of text data are often ill-structured, contain many self-invented acronyms, shorthand and typos We developed two text data mining systems successfully being deployed in Ford/Volvo

Noon to 12:30 p.m. **Technical Presentations (1), 1670 CSE**

Lujun Fang, Kristen LeFevre, CSE
Privacy Wizards for Social Networking Sites

Privacy is an enormous problem in online social networking sites. While sites such as Facebook allow users fine-grained control over who can see their profiles, it is difficult for average users to specify this kind of detailed policy. In this paper, we propose a template for the design of a social networking privacy wizard. The intuition for the design comes from the observation that real users conceive their privacy preferences (which friends should be able to see which information) based on an implicit set of rules. Thus, with a limited amount of user input, it is usually possible to build a machine learning model that concisely describes a particular user's preferences, and then use this model to configure the user's privacy settings automatically. As an instance of this general framework, we have built a wizard based on an active learning paradigm called uncertainty sampling. The wizard iteratively asks the user to assign privacy “labels” to selected (“informative”) friends, and it uses this input to construct a classifier, which can in turn be used to automatically assign privileges to the rest of the user's (unlabeled) friends. To evaluate our approach, we collected detailed privacy preference data from 45 real Facebook users. Our study revealed two important things. First, real users tend to conceive their privacy preferences in terms of communities, which can easily be extracted from a social network graph using existing techniques. Second, our active learning wizard, using communities as features, is able to recommend high-accuracy privacy settings using less user input than existing policy-specification tools.

Ahmet Duran, Assistant Professor, Mathematics
Daily return discovery in financial markets

I study large data sets of daily market price returns for S&P 500-listed stocks and closed-end funds (CEFs) trading in US markets from 1998 to 2009 using new data mining approaches. I believe that a complex combination of motivations leads price dynamics. These motivations may contain multi-actions such as active, reactive, interactive, proactive and delayed decision making components where one of them may dominate at a time depending on short or long term considerations. I find a statistically significant combined pattern of delayed-reaction and overreaction.

Yongqun “Oliver” He, Medical School
(Lab Overview)

The He Research Group (http://www.hegroup.org) conducts both wet-lab and dry-lab (bioinformatics) biomedical research. Our wet-lab research focuses on understanding Brucella-host interaction mechanisms and develop effective Brucella vaccines. Our bioinformatics research aims to develop database and analysis systems to study host-pathogen interactions and vaccines against infectious diseases. Our bioinformatics approaches includes development of web-based databases, ontology systems, Bayesian network-based microarray analysis algorithms, and literature tools. These systems biology approaches are being explored to study Brucella and other infectious pathogens, biological interaction networks, or other biomedical problems.

12:30 to 1:30 p.m. **Lunch: TISHMAN HALL, CSE**

1:30 to 2:45 p.m. **SESSION III, Technical Presentations (2), 1670 CSE**

Jungkap Park, Mechanical Engineering, Gus R. Rosania, Pharmaceutical Sciences, and Kazuhiro Saitou, Mechanical Engineering
Tunable Machine Vision-Based Strategy for Automated Annotation of Chemical Databases

We present a tunable, machine vision-based strategy for automated annotation of virtual small molecule databases. The proposed strategy is based on the use of a machine vision-based tool for extracting structure diagrams in research articles and converting them into connection tables, a virtual "Chemical Expert" system for screening the converted structures based on the adjustable levels of estimated conversion accuracy, and a fragment-based measure for calculating intermolecular similarity. For annotation, calculated chemical similarity between the converted structures and entries in a virtual small molecule database is used to establish the links. The overall annotation performances can be tuned by adjusting the cutoff threshold of the estimated conversion accuracy. We perform an annotation test which attempts to link 121 journal articles registered in PubMed to entries in PubChem which is the largest, publicly accessible chemical database. Two cases of tests are performed, and their results are compared to see how the overall annotation performances are affected by the different threshold levels of the estimated accuracy of the converted structure. Our work demonstrates that over 45% of the articles could have true positive links to entries in the PubChem database with promising recall and precision rates in both tests. Furthermore, we illustrate that the Chemical Expert system which can screen converted structures based on the adjustable levels of estimated conversion accuracy is a key factor impacting the overall annotation performance. We propose that this machine vision-based strategy can be incorporated with the text-mining approach to facilitate extraction of contextual scientific knowledge about a chemical structure, from the scientific literature.

Arnab Nandi, H.V. Jagadish, CSE
Autocompletion for Structured Querying

Given the increasing complexity and size of data accessed by the average end user, the construction of rich, meaningful queries has become a challenge. An ideal query interface enables the user to easily construct such queries and allows the user to explore the data without expert knowledge of the system. We describe the adaptation of autocompletion, a familiar and interactive interface mechanism, as a querying method that meets these demands. We identify the challenges in building an autocompletion-based query interface: efficiently guiding the user through the query space, ensuring query correctness and concisely representing structured information to the user. We present solutions to these challenges by using both the structure and instances of the data while interacting with the user. We demonstrate that the use of autocompletion is both a viable and convenient solution to structured query construction.

Christopher J. Miller, Astronomy
Astronomy in the Cloud: The Virtual Observatory

The fields of astronomy and astrophysics have embarked on a journey to connect the world's astronomical data archives. Each of the datasets within these archives is unique and they share only a few common attributes amongst each other. Each archive was designed to meet specific scientific goals and thus their schemas, interfaces, and access mechanisms are heterogeneous. Yet when these data are properly combined, we are able to achieve new and important scientific insight. Thus, the challenge to create a truly Virtual Observatory is an important one. I will provide an overview of the types of astronomical data and archives which exist today and in the future and discuss the middleware standards and technologies that are being developed to move astronomy off the desktop and into the cloud.

Matthew Brook O'Donnell and Nick C. Ellis, Linguistics

Extracting an Inventory of English Verb Constructions from Language Corpora

This paper outlines and pilots our approach towards developing an inventory of verb argument constructions based upon English form, function, and usage. We search a tagged and dependency-parsed corpus (the BNC - a 100-million word corpus of English) for Verb-Argument Constructions (VACs) including those previously identified in the pattern grammar resulting from the COBUILD project. This generates: (1) a list of verb types that occupy each construction (2) a frequency ranked type-token distribution for these verbs, and we determine the degree to which this is Zipfian (3) a contingency-weighted list which reflects their statistical association (faithfulness). We believe that each of these measures is a step towards increasing the learnability of VACs as categories following principles of associative learning. One test of this is whether there is an increase across lists of the semantic cohesion of the verbs occupying each construction, and whether they follow a prototype/radial category structure. From inspection, this seems to be so. We are developing measures of this using network measures of clustering in the verb-space defined by WordNet and Roget's Thesaurus. Future developments will explore the use of supervised learning methods to train classifiers that can be used to mine VACs from less richly annotated corpora.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu, Statistics

Joint Estimation of Multiple Graphical Models

Gaussian graphical models explore dependence relationships between random variables, through estimation of the corresponding inverse covariance matrices. The objective of this paper is to develop an estimator for such models appropriate for heterogeneous data. In this setting, estimating a single graphical model would mask the underlying heterogeneity, while estimating separate graphical models for each category ignores the common structure. We propose a method which jointly estimates several graphical models corresponding to the different categories present in the data. The method aims to preserve the common structure, while allowing for differences between the categories. This is achieved through a hierarchical penalty that targets the removal of common zeros in the inverse covariance matrices across categories. We establish the asymptotic consistency and sparsistency of the proposed estimator in the high-dimensional case, and illustrate its superior performance on a number of simulated networks. An application to learning semantic connections between terms from web pages collected from computer science departments is also included.

Ahmed Hassan, CSE, Rosie Jones, Yahoo! Labs, and Kristina Klinkner, Carnegie-Mellon University

Beyond DCG: User Behavior as a Predictor of a Successful Search

Web search engines are traditionally evaluated in terms of the relevance of web pages to individual queries. However, relevance of web pages does not tell the complete picture, since an individual query may represent only a piece of the user's information need and users may have different information needs underlying the same queries. We address the problem of predicting user search goal success by modeling user behavior. We show empirically that user behavior alone can give an accurate picture of the success of the user's web search goals, without considering the relevance of the documents displayed. In fact, our experiments show that models using user behavior are more predictive of goal success than those using document relevance. We build novel sequence models incorporating time distributions for this task and our experiments show that the sequence and time distribution models are more accurate than static models based on user behavior, or predictions based on document relevance.

2:45 to 3:30 p.m. **Panel Discussion, 1670 CSE**

"Do we have the critical mass to achieve something outstanding together"
Dragomir Radev, Moderator
H.V. Jagadish, CSE
Michael Cafarella, CSE
Kristen LeFevre, CSE
Fan Meng, Microarray Laboratory, MBNI
Chris Miller, Astronomy
Raghu Ramakrishnan

3:30 to 4:00 p.m. **Poster Session, 2nd Floor CSE**

Daniel Fabbri, Kristen LeFevre and Qiang Zhu, CSE
PolicyReplay: Misconfiguration-Response Queries for Data Breach Reporting

Recent legislation has increased the requirements of organizations to report data breaches, or unauthorized access to data. While access control policies are used to restrict access to a database, these policies are complex and difficult to configure. As a result, access control policies are sometimes misconfigured, allowing users to inappropriately read or modify data. In this paper, we consider the problem of reporting data breaches after such a misconfiguration is detected. A strawman solution to this problem is to go back in time and replay all the operations that occurred in the interim, with the correct policies, to determine the set of queries that were affected by the misconfiguration, which is inefficient. Instead, we develop techniques to improve reporting efficiency by reducing the number of operations that must be re-executed and reducing the cost of replaying the operations. An extensive evaluation shows that our method can reduce the total runtime by up to an order of magnitude.

Jiang Yang and Scott Counts, SI
Predicting the Speed, Scale, and Range of Information Diffusion in Twitter

We present results of network analyses of information diffusion on Twitter, via users' ongoing social interactions as denoted by "username" mentions. Incorporating survival analysis, we constructed a novel model to capture the three major properties of information diffusion: speed, scale, and range. On the whole, we find that some properties of the tweets themselves predict greater information propagation but that properties of the users, the rate with which a user is mentioned historically in particular, are equal or stronger predictors.

Matthew Simmons, Gayathri Balasubramanian, Eytan Adar, and Lada Adamic, SI
Text mutation in online diffusion

One unexplored area of information diffusion is an examination of the changes, or mutations, introduced into information as it diffuses. In this poster we explore mimetic mutation in information diffusion by examining changes in quotations, hyperlinks, and their surrounding context in online news articles and blog posts.

Pradeep Muthukrishnan, Amjad abu Jbara, Vahed Qazvinian, Dragomir R. Radev, CSE

Citation summarization of scientific literature

We introduce the ACL Anthology Network (AAN) a manually curated database of publications, citations and collaborations in the field of Computational Linguistics. AAN consists of 15,160 papers with 62,825 citations between them. This data set has already been used in a variety of experiments including summarization, reference extraction, topic modeling, etc. In this work we present the first steps in producing an automatically generated, readily consumable, technical survey. Specifically we explore the combination of citation information and summarization techniques and show that in the framework of multi-document survey creation, citation texts can play a crucial role.

Xiao Wei, Jiang Yang, Lada A. Adamic, SI

Viral diffusion of social games on Facebook

Social games, being embedded in and drawing upon existing social networks, have the potential to spread virally. In this paper, we examine two popular social games, each having millions of Facebook users, to understand the role that users play individually and collectively in propagating social applications. At the individual level, the users' invitation behavior significantly outweighs their demographic and social network properties in predicting invitation success rate. At the collective level, we demonstrate that social games that encourage group formation tend to rapidly integrate dense network cliques. Finally, engagement in a social game is closely tied with the ability to recruit friends.

Eytan Bakshy, Matthew P. Simmons, David A. Huffaker, Chun-Yuen Teng, Lada A. Adamic, SI

The Social Dynamics of Economic Activity in a Virtual World

This paper examines social structures underlying economic activity in Second Life (SL), a massively multiplayer virtual world that allows users to create and trade virtual objects and commodities. We find that users conduct many of their transactions both within their social networks and within groups. Using frequency of chat as a proxy of tie strength, we observe that free items are more likely to be exchanged as the strength of the tie increases. Social ties particularly play a significant role in paid transactions for sellers with a moderately sized customer base. We further find that sellers enjoying repeat business are likely to be selling to niche markets, because their customers tend to be contained in a smaller number of groups. But while social structure and interaction can help explain a seller's revenues and repeat business, they provide little information in the forecasting a seller's future performance. Our quantitative analysis is complemented by a novel method of visualizing the transaction activity of a seller, including revenue, customer base growth, and repeat business.

Chun-Yuen Teng and Lada Adamic, SI

Longevity in Second Life

User retention is important to the success of online social media, particularly in virtual world settings where users shape one another's online experience. We study a rich set of variables, including social network and group membership, chatting, and transactions, in order to predict which users will stay and which ones will leave. We find that simple variables directly measuring the intensity and diversity of a user's interaction with others are most predictive.

Ahmed Hassan and Dragomir R. Radev, CSE

Computational Sociolinguistics

There is a massive amount of text produced daily on the Web. Search engines are playing a vital role in making this information accessible to users. However, the rise of Web 2.0 technologies brought in several new research challenges that go beyond search. One such challenge is combining social and linguistic analysis to unveil relations in social media. Computational sociolinguistics is a new scientific field, at the intersection of natural language processing, and social network analysis. Its central challenge is to use linguistics analysis techniques to understand social relations that develop in virtual Web communities by analyzing user generated textual content. In this work, we present several methods in computational sociolinguistics that exploit the synergies between text and networks to discover social relations from Web text information. These methods are used for analyzing the blogosphere, tracking influence in political speeches, and mining polarized discussions on the Web.

Chun-Yuen Teng, Debra Lauterbach, and Lada A. Adamic, SI

I rate you. You rate me. Should we do so publicly?

We find that ratings are not absolute, but rather depend on whether they are given anonymously or under one's own name and whether they are displayed publicly or held confidentially. The potential to reciprocate produces higher and more correlated ratings than when individuals are unable to see how others rated them. Ratings further depend on the gender and nationalities of the raters and ratees. All of these findings indicate that ratings should not be taken at face value without considering social nuances.

W. Xuan; M. Dai; J. Buckner; B. Mirel; J. Song; H. Dong; M. Bota; B. Athey; H. Akil; W. Stanley; F. Meng, Microarray Laboratory, MBNI

Integrated Literature and Data Exploration Using Interactive Brain Maps

Existing literature exploration solutions treat anatomical structures as concepts or at most concept diagrams, missing critical spatial information about different structures such as size, location, substructure and the relationships among them. Such information is important for developing comprehensive understanding of the biological implications of structure-specific data such as brain imaging data and gene expression patterns in different brain regions. PubAnatomy is our attempt to use interactive anatomical structure graphs derived from voxel level imaging data to address some of the issues for literature-driven exploration of neurobiological data sets. It provides new ways to explore relationships among brain structures, pathophysiological processes, gene expression levels and protein-protein interactions by presenting Medline literature and experimental data in the context of mouse brain anatomy and gene network. Our solution can be easily extended to any annotated voxel level imaging data as well as integrating third party molecular level data and analysis functions. The prototype of PubAnatomy is at: http://brainarray.mbni.med.umich.edu/Brainarray/prototype/PubAnatomy.

Arzucan Ozgur, Dragomir R. Radev, Amjad abu Jbara, CSE

Information Extraction from Biomedical Text

One of the greatest challenges that the researches in the Biomedical domain face is that, most of the knowledge remains hidden in the unstructured text of the large number of published articles. Developing text mining methods to automatically uncover this hidden knowledge is not only useful, but also necessary to facilitate biomedical research. We investigate methods to automatically extract information about protein and gene interactions contained in the biomedical scientific literature. Besides extracting the fact that two entities interact, we also address the problem of extracting the interaction context such as type, directionality, and certainty. We concentrate on taking one step further by integrating information extraction with network analysis techniques to infer novel (potentially currently unknown) relationships between biomedical entities.

Manhong Dai, Nigam H. Shah, Weijian Xuan, Mark A. Musen, Stanley J. Watson, Brian Athey and Fan Meng, Microarray Laboratory, MBNI

An Efficient Solution for Mapping Free Text to Unique Biomedical Concepts

Full and semi-ontology systems such as the Open Biomedical Ontologies and the Unified Medical Language System provide the basis for automated cross-domain data integration and inference. A major challenge for using such ontology systems is the mapping of free text terms to ontology terms. Here we present an efficient solution that is about two orders of magnitude faster and more flexible than the popular MMTx program. Our solution will enable the on-the-fly mapping of biomedical literature and clinical records to ontology terms.

Kevin S. Xu, EECS, Mark Kliger, Medasense Biometrics Ltd., Israel, Alfred O. Hero III, EECS

Identifying Spammers by Their Resource Usage Patterns

Most studies on spam thus far have focused on its content or source. These types of studies, however, reveal little about the behavioral characteristics of spammers. In addition, privacy issues may prevent wide access to email content. In this paper, we try to identify spammers by investigating their resource usage patterns. Specifically, we look at usage patterns of harvesters, the bots that are used to acquire email addresses, and spam servers, the email servers being used to send the spam emails. We perform spectral biclustering on both harvesters and servers to reveal groups of resources that are used together, which we believe correspond to individual spammers or groups of spammers. We make several interesting discoveries including a division into phishing and non-phishing spammers and a group of harvesters with highly correlated behavior that have IP addresses belonging to a rogue Internet service provider.

Yunpeng Zhao, Elizaveta Levina and Ji Zhu, Statistics

Extracting communities from networks

Analysis of networks and in particular discovering communities within networks has been a focus of recent work in several fields, with applications ranging from citation and friendship networks to food webs and gene regulatory networks. Most of the existing community detection methods focus on partitioning the network into cohesive communities, with the expectation of many links between the members of the same community and few links between different communities. However, many real-world networks contain, in addition to communities, a number of sparsely connected nodes that are best classified as “background”. To address this problem, we propose a new criterion for community extraction, which aims to separate tightly linked communities from a sparsely connected background, extracting one community at a time. The new criterion is shown to perform well in simulation studies and on several real networks. We also establish asymptotic consistency of the proposed method under the block model assumption.

Xiaodan Zhou, Paul Resnick, SI

Extracting the Meaning of Political Concepts in Chinese Online Discussion

Most political concepts have divergent semantic meanings in different settings: for example, the word “democracy” certainly means differently in the communist propaganda than in the liberal discourse. This research proposal applies text mining techniques and extracts, explains and compares the meanings of "democracy" in 4 different settings: 1) the Chinese official propaganda, 2) heavily-regulated Chinese online

political discussion, 3) lightly-regulated Chinese online political discussion, and 4) US political blogs. The technical contribution is to propose a novel way of constructing concepts network using the “random walk with restart” algorithm. In addition, the study proposes an evaluation framework to optimize the construction of semantic representation model from text corpuses. The domain specific contribution is to use computational techniques to provide empirical evidence for political and social sciences research in terms of understanding the meaning of “democracy” in the Chinese context.

4:00 to 5:00 p.m. **CSE Distinguished Lecture**: ***“Building and Searching a Web of Concepts,” Raghu Ramakrishnan, Yahoo! 1*670 CSE**

Search engines are increasingly offering results that are based on a semantically rich interpretation of the user's intent and the content available to satisfy that intent. A natural question is to ask how far along we are in understanding content on the web. The Semantic Web seeks to enable publication of data with rich markups that facilitate automated interpretation; Yahoo!'s Search Monkey is an example of a service in this spirit. However, there is much useful data that is not semantically marked up, and many domains in which the coverage of existing structured data feeds is low. In this talk, I will discuss the goal of constructing a web of "concepts" (a term I use to denote entities, categories of entities, and relationships) by starting with the current view of the web (as a collection of hyperlinked pages, or documents, each seen as a bag of words).
We need to extract concept-centric metadata for a broad and deep set of important concepts, and stitch it together to create a semantically rich aggregate view of all the information available on the web for each concept instance. The goal of building and maintaining such a web of concepts presents many challenges, but also offers the promise of enabling many powerful applications, including novel search and information discovery paradigms. In this talk, I will describe a research agenda towards this goal and discuss related work, including the PSOX project at Yahoo!.

Raghu Ramakrishnan’s Biography

Raghu Ramakrishnan is Chief Scientist for Audience and Cloud Computing at Yahoo!, and a Yahoo! Fellow, Building and Searching a Web of Concepts. His work has influenced query optimization in commercial database systems and the design of window functions in SQL:1999. His paper on the Birch clustering algorithm received the SIGMOD 10-Year Test-of-Time award, and he has written the widely-used text "Database Management Systems" (with Johannes Gehrke). Ramakrishnan is a Fellow of the ACM and IEEE, and has received several awards, including the ACM SIGKDD Innovations Award, the ACM SIGMOD Contributions Award, a Distinguished Alumnus Award from IIT Madras, a Packard Foundation Fellowship in Science and Engineering, and an NSF Presidential Young Investigator Award. He is Chair of ACM SIGMOD, on the Board of Directors of ACM SIGKDD and the Board of Trustees of the VLDB Endowment. Ramakrishnan was Professor of Computer Sciences at the University of Wisconsin-Madison, and founder and CTO of QUIQ, a company that pioneered question-answering communities, powering Ask Jeeves' AnswerPoint as well as customer-support for companies such as Compaq. Raghu Ramakrishnan got his B.Tech. from IIT Madras in 1983 and his Ph.D. from the University of Texas at Austin in 1987.

5:00 to 5:10 p.m. **Q&A Session, follows the Distinguished Lecture, 1670 CSE**

5:15 to 6:00 p.m. **DISTINGUISHED LECTURE RECEPTION, 3rd floor lounge area of CSE and POSTER SESSION CONTINUES, 2nd Floor, CSE**