MITACS Winter School on
Modelling and Mining of Networked Information Spaces

BIRS, Banff, Alberta

November 25 - 28, 2006
`http://www.mathstat.dal.ca/~waw2006/winterschool`

A library, no matter how vast, is limited by the efficiency of its cataloguing and searching system. So it is with the World Wide Web. With the emergence of the Web as the pre-eminent information storage and retrieval system, the need for better organization and searching systems has become pressing. Between the pin-prick of desire for information and the fruition of a screen full of relevant links lies the latency of intelligent web search. We are getting used to seeing the world through the Web; the time has come to see every world as a web. Every collection of interlinked data (a networked information space) can be organized and searched like the Web. Networks of phone calls, networks of financial transactions, networks of social interactions can all be analyzed by the techniques used for the Web. The secret to successfully searching the web has been known to the builders of better search engines. On the web, as in other networked information spaces, the required information about the information lies in the link structure. Understanding this link structure mathematically and being able to translate this knowledge into applications is the point of this project. The theorems, tools and techniques of graph theory are at our disposal.

The MoMiNIS Winter School provides the opportunity to a group of selected graduate students to attend tutorials in the Modelling and Mining of Networked Information Spaces from recognized experts in the field, and present their own work and receive feedback. The unique setting of BIRS is highly conducive to personal interaction and generation of new ideas. The Winter School has the following objectives:
To provide knowledge of the state of the art in the field via tutorials by experts.
To support young researchers with networking.
To provide young researchers with feedback on their research and the opportunity to discuss their research interests with established researchers in the field.

The Winter School is most useful to students who are starting their doctoral research and have some ideas about the direction of their PhD thesis.

We gratefully acknowledge the financial support from:

The Mathematics of Information Technology and Complex systems (MITACS)
Network of Centres of Excellence,
Burnaby, British Columbia, Canada
`www.mitacs.math.ca`

Yahoo! Inc.,
Sunnyvale, California,
`www.yahoo.com`

Genieknows.com
Halifax, Nova Scotia
`www.genieknows.com`

MITACS Winter School
on
Modelling and Mining of Networked Information Spaces
BIRS event 06w5104
Schedule

	Saturday, November 25
16:00	Check-in begins (Front Desk Professional Development Centre - open 24 hours)
17:30-19:30	Buffet Dinner, Donald Cameron Hall
20:00	Informal gathering in 2nd floor lounge, Corbett Hall Opportunity to prepare poster display

	Sunday, November 26
8:30	Opportunity for students to set up posters
9.15	Introduction and Welcome to BIRS by BIRS Station Manager
9:30	Welcome and introduction
10.15	Coffee break
10:30	**Tutorial**: Walter Willinger *On measuring, inferring, and modeling Internet connectivity: A guided tour across the TCP/IP protocol stack [p. 4]*
11:30	Lunch (Donald Cameron Hall)
13:00	Campus tour
14:00	Discussion in groups (Max Bell auditorium)
14:30	Walter Willinger (continued)
15:30	Coffee break
15:45	**Poster presentations** (WW and SC group) Angelova, Ralitsa (SC) [p. 8] Meiss, Mark (WW) [p. 10] Pandey, Guaray (WW) [p. 11] Shafiei, Mahdi (SC) [p. 11] Tanta-ngai, Hathai (SC) [p. 12] Thomas, Dilys (WW) [p. 12] Wang, Tao (SC) [p. 13] Zhou, Bin (WW) [p. 13]
17:30-19:30	Dinner (Donald Cameron Hall)

	Monday, November 27
9:00	**Tutorial**: Filippo Menczer
	Web Mining, mapping, modeling and mingling
10:00	**Poster presentations** (FM group) [p. 5]
	Awekar, Amit (FM) [p. 8]
	Barouni Ebrahimi, Mohammadreza (FM) [p. 8]
	Gurevich, Maxim (FM) [p. 8]
	Kasneci, Gjergji (FM) [p. 9]
10:45	Coffee break
11:00	Filippo Menczer (continued)
12:00	Lunch (Donald Cameron Hall)
13:30	**Tutorial**: Soumen Chakrabarti
	Ranking and labeling graphs:
	Analysis of links and node attributes [p. 5]
15:30	Coffee break and problem session
16:00	Soumen Chakrabarti (continued)
17:30-19.30	Dinner (Donald Cameron Hall)

	Tuesday, November 28
9:00	**Tutorial**: David Liben-Nowell
	Navigation and Evolution of Social Networks [p. 7]
10:00	**Poster presentations** (LN and FC group)
	Healy, John (LN) [p. 9]
	Horn, Paul (FC) [p. 9]
	Imani, Navid (FC) [p. 9]
	McKay, Neil (FC) [p. 10]
	Nargis, Isheeta (FC) [p. 10]
	Olsen, Martin (LN) [p. 10]
	Vlachou, Akrivi (LN) [p. 12]
	Wan, Xiaomeng (LN) [p. 12]
10:45	Coffee break
11:00	David Liben-Nowell (continued)
12:00	Lunch (Donald Cameron Hall)
13:20	Group Photo
13:30	**Tutorial**: Fan Chung-Graham
	Graph Theory in the New Millennium [p. 7]
14:30	Problem solving session
15:15	Fan Chung-Graham (discussion)
15:45	Coffee break
16:00	**Panel discussion**
17:30-19.30	Dinner (Donald Cameron Hall)

On measuring, inferring, and modeling Internet connectivity: A guided tour across the TCP/IP protocol stack

Walter Willinger
AT&T Labs-Research

One of the most visible manifestations of the Internet's vertical decomposition is the 5-layer TCP/IP protocol stack. This layered architecture gives rise to a number of different connectivity structures, with the lower layers (e.g., router-level) defining more physical and the higher layers (e.g., the Web) more virtual or logical types of topologies. The resulting graph structures have been designed with very different objectives in mind, have evolved according to different circumstances, and have been shaped by succinctly different forces. The main objective of this tutorial is to discuss the problems and challenges associated with measuring, inferring, and modeling these different connectivity structures. To this end, the tutorial is divided into the following four parts:

(1) Measurements: Internet connectivity measurements are notorious for their ambiguities, inaccuracies, and incompleteness. As a general rule, they should never be taken at face value, but need to be scrutinized for consistency with the networking context from which they were obtained, and to do so, it is important to understand the process by which they were collected.

(2) Inference: The challenge is to know whether or not the results we infer from our measurements are indeed well-justified claims, and at issue are the quality of the measurements themselves, the quality of their analysis, and the sensitivity of the inferred properties to known imperfections of the measurements.

(3) Modeling: Developing appropriate models of Internet connectivity that elucidate observed structure or behavior is typically an underconstrained problem, meaning that there are in general many different explanations for one and the same phenomenon. To argue in favor of any particular explanation typically involves additional information, either in the form of domain knowledge or of new or complementary data. It is in the choice of this side information and how it is incorporated into the model building process, where considerable differences arise in the various approaches to Internet topology modeling that have been applied to date.

(4) Model validation: There has been an increasing awareness of the fact that the ability to replicate some statistics of the original data or inferred quantities does not constitute validation for a particular model. While one can always use a model with enough parameters to "fit" a given data set, such models are merely descriptive and have in general no explanatory power. For the problems described here, appropriate validation typically means additional work (e.g., identifying and collecting complementary measurements that can be used to check a proposed explanation).

The tutorial requires some basic understanding of the Internet architecture and of existing Internet technologies, and knowledge of basic concepts from mathematics, statistics, and graph theory will be helpful. There will be ample opportunities to ask questions, explore particular problems, and discuss alternative perspectives.

Suggested reading materials:

- L. Li, D. Alderson, W. Willinger, and J. Doyle, "A first-principles approach to understanding the Internet's router-level topology"

  Proc. ACM SIGCOMM 2004, pp. 3-14 (2004). `http://www.sigcomm.org/sigcomm2004/papers/p599-li.pdf`

- J. Doyle, D. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger, "The 'robust yet fragile' nature of the Internet" Proc. Nat. Acad. Sci. 102(41):14497-14502 (2005). `http://www.pnas.org/cgi/reprint/102/41/14497`

- D. Alderson, H. Chang, M. Roughan, S. Uhlig, and W. Willinger, "The many facets of Internet topology and traffic" Networks and Heterogeneous Media 1(4):569-600 (2006). `http://aimsciences.org/journals/NHM/nhm_online.jsp`

- H. Chang, S. Jamin, and W. Willinger, "To peer or not to peer: Modeling the evolution of the Internet's AS-level topology"

  Proc. IEEE INFOCOM (2006). `http://topology.eecs.umich.edu/archive/infocom06.pdf`

- D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, "On unbiased sampling for unstructured peer-to-peer networks" Proc. ACM/USENIX Internet Measurement Conference (IMC'06) (2006). `http://www.barsoom.org/~agthorr/papers/imc-2006-sampling.pdf`

Walter Willinger received the Diplom (Dipl. Math.) from the ETH Zurich, Switzerland, and the M.S. and Ph.D. degrees from the School of ORIE, Cornell University, Ithaca, NY. He is currently a member of the Information and Software Systems Research Center at AT&T Labs - Research, Florham Park, NJ, and before that, he was a Member of Technical Staff at Bellcore Applied Research (1986-1996). His research interests include studying the multiscale nature of Internet traffic and topology and developing a theoretical foundation for dealing with large-scale communication networks such as the Internet. He is a Fellow of ACM (2005) and a Fellow of IEEE (2005). For his work on the self-similar ("fractal") nature of Internet traffic, he received the 1996 IEEE W.R.G. Baker Prize Award, the 1994 W.R. Bennett Prize Paper Award, and the 2005 ACM/SIGCOMM "Test of Time" Paper Award.
Email address: walter@research.att.com.

Web Mining, mapping, modeling and mingling

Filippo Menczer
Indiana State University

The Web is a complex self-organized system whose evolution and use is shaped by many concurrent social, cognitive, economic, and information phenomena. This tutorial will describe ongoing efforts to study the topological and dynamical properties of link, content, and semantic networks stemming from some of these forces. We will discuss what we think, what we know, what we can use regarding the structure, content, and use of the Web, and what the future of intelligent, cooperative Web search may bring.

Filippo Menczer is an associate professor of informatics and computer science, adjunct associate professor of physics, and a member of the cognitive science program at Indiana University, Bloomington. He holds a Laurea in Physics from the University of Rome and a Ph.D. in Computer Science and Cognitive Science from the University of California, San Diego. Dr. Menczer has been the recipient of Fulbright, Rotary Foundation, and NATO fellowships, and is a fellow-at-large of the Santa Fe Institute. His research is supported by a Career Award from the National Science Foundationon and focuses on Web, text, and data mining, Web intelligence, distributed information systems, social Web search, adaptive intelligent agents, complex systems and networks, and artificial life.

Ranking and labeling graphs: Analysis of links and node attributes

Soumen Chakrabarti
IIT Bombay

We will study techniques for ranking and labeling nodes in a graph, based on the link structure of the graph as well as attributes of the nodes. Ranking and labeling have obvious applications in Web search and page classification, but the range of applications is widening to finer-grained entity-relationship graphs where nodes represent entities like people, emails, papers, organizations and locations and edges represent relations like works-for, wrote, cited, is-located-in and so on. Applications also include annotating unstructured and semistructured sources with type tags which can then be indexed for search.

On the subject of ranking, we will cover the following topics (two hours):

- Hyperlink induced topic search (HITS); identifying the relevant subgraph, connections with SVD/PCA and term-document random walks and translation models
- Pagerank; dead ends and teleport, choice of decay profiles, large-scale computational issues
- Comparison between HITS and Pagerank; scoring and systems issues, content focus and topic drift, topological sensitivity, score and rank stability
- Probabilistic variants of HITS: SALSA, PHITS
- Pagerank variants; personalized Pagerank, absorbing random walks and applications to page staleness detection, link spam detection via trust and mistrust propagation, viral marketing, graph-based similarity and graph fingerprints
- Adding content to HITS: vector-space model, anchor text, HTML tag-tree
- Adding content to Pagerank: word proximity search, the "intelligent surfer", topic-sensitive Pagerank, ObjectRank
- Learning to rank: learning a weight vector for dot-product scoring and ranking

- Extending to Markov walks and the conductance matrix
- Maximum entropy network flows: spotting a hidden favored community
- Edge typed and weights in the conductance matrix

On the subject of labeling, we will cover the following topics (1.5 hours):

- Examples of network influence: Web page topics, social networks and purchase habits
- Applications: page classification, text tagging, entity resolution, viral marketing
- Markov fields, cliques, potential functions, weights
- Inferencing and model estimation
- Easy cases: chains and trees
- General graphs: relaxation labeling
- Associative networks: linear and quadratic programming relaxations
- Max-margin transudctive labeling

Only basic calculus and matrix algebra will be assumed. There might be some small experiments using Scilab and Java and small-scale data sets.

Soumen Chakrabarti received his B.Tech in Computer Science from the Indian Institute of Technology, Kharagpur, in 1991 and his M.S. and Ph.D. in Computer Science from the University of California, Berkeley in 1992 and 1996. At Berkeley he worked on compilers and runtime systems for running scalable parallel scientific software on message passing multiprocessors.
He was a Research Staff Member at IBM Almaden Research Center from 1996 to 1999, where he worked on the Clever Web search project and led the Focused Crawling project.
In 1999 he joined the Department of Computer Science and Engineering at the Indian Institute of Technology, Bombay, where he has been an Associate professor since 2003. In Spring 2004 he was Visiting Associate professor at Carnegie-Mellon University.
He has published in the WWW, SIGIR, SIGKDD, SIGMOD, VLDB, ICDE, SODA, STOC, SPAA and other conferences as well as Scientific American, IEEE Computer, VLDB and other journals. He holds eight US patents on Web-related inventions. He has served as technical advisor to search companies and vice-chair or program committee member for WWW, SIGIR, SIGKDD, VLDB, ICDE, SODA and other conferences, and guest editor or editorial board member for DMKD and TKDE journals. He is also author of a book on Web Mining.
His current research interests include integrating, searching, and mining text and graph data models, exploiting types and relations in search, and Web graph and popularity analysis.

Navigation and Evolution of Social Networks

David Liben-Nowell
Carleton College

Imagine yourself in Ancient Greece, newly crowned as a chariot-racing champion, when you are accidentally stabbed by the laurel wreath with which your achievement is being marked. But is laurel poisonous? What is the antidote? 2400 years before the invention of the web, you would have to seek this information from your friends and, indirectly, their friends, their friends' friends, and so on. Social networks are formal structures that encode these connections between people, and the recent explosion of online data recording social interactions has granted us the opportunity to analyze these networks with the full power of algorithmic graph theory. In this tutorial, I will introduce some of the empirical observations of the structure of social networks, especially in comparison to the structure of the web. We will then discuss a number of algorithmic topics arising in social networks, including the latent information contained in social networks (how much information about people is implicit in their connections?) and how to search social networks (can you find a short path to a target without global knowledge of the graph?).

David Liben-Nowell is an assistant professor of computer science at Carleton College, which is located in the part of Minnesota that makes fun of Duluth for having tough winters. He received his PhD in theoretical computer science from MIT's Computer Science and Artificial Intelligence Laboratory in 2005. His research interests include a variety of applications of the techniques of theoretical computer science to questions arising within and beyond computer science, with a focus on large-scale information networks and their evolution. David's research interests also include game theory, peer-to-peer computing, and computational biology. Prior to coming to MIT, David received a BA from Cornell and an MPhil from the University of Cambridge. While living in the UK, he developed an unhealthy obsession with cricket; he also enjoys ultimate frisbee, road biking, and constructing crossword puzzles in his free time.

Graph Theory in the Information Age

Fan Chung-Graham
University of California, San Diego

We will discuss some basic methods in graph theory and random graphs that are useful for analyzing large information networks.

Fan Chung-Graham received a B.S. degree in mathematics from National Taiwan University in 1970 and a Ph.D. in mathematics from the University of Pennsylvania in 1974, after which she joined the technical staff of AT&T Bell Laboratories. From 1983 to 1991, she headed the Mathematics, Information Sciences and Operations Research Division at Bellcore. In 1991 she became a Bellcore Fellow. In 1993, she was the Class of 1965 Professor of Mathematics at the the University of Pennsylvania. Since 1998, she has been a Professor of Mathematics and Professor of Computer Science and Enginering at the University of California, San Diego. She is also the Akamai Professor in Internet Mathematics.

Her research interests are primarily in graph theory, combinatorics, and algorithmic design, in particular in spectral graph theory, extremal graphs, graph labeling, graph decompositions, random graphs, graph algorithms, parallel structures and various applications of graph theory in Internet computing, communication networks, software reliability, chemistry, engineering, and various areas of mathematics. She was awarded the Allendoerfer Award by Mathematical Association of America in 1990. Since 1998, she has been a member of the American Academy of Arts and Sciences.

Poster abstracts

(Ordered alphabetically by presenter's last name)

Neighborhood Watch: Document Classification in Typed Graphs

Ralitsa Angelova
Max-Planck Institute for Informatics

Classification is a challenging problem with a broad impact on areas like machine learning, information retrieval, pattern recognition, image analysis and bioinformatics. Recent research shows that incorporating relationships into the classification process is beneficial but poses difficulties which, if carelessly addressed, degrade the classification result. One hard problem is how to make use of the link structure in a heterogeneous environment. Such an environment is represented by a graph containing nodes of different types. Nodes of the same type as well as nodes that belong to different types are connected by edges (links). Different node types have different systems of possible class labels. The goal is to assign to each node the best suitable label among its possible classes according to a local likelihood and the node's neighborhood in the graph. We propose a relaxation labeling approach for classification of heterogeneous graphs.
This is joint work with Prof. Gerhard Weikum.

Link Analysis Based Methods for Handling Abundance and Misrepresentation Over The Web

Amit Awekar
North Carolina State University

Broad aim of our work is to investigate how link analysis based methods can be useful to deal with abundance and misrepresentation issues over the Web. We have worked over an algorithm SelHITS, for answering broad-topic queries over the Web. We want to apply same approach to other topic oriented tasks over the Web. Clustering hypertext repository is our current problem of interest. Our approach is to iteratively modify the representation of documents using link based ranking functions.

Towards Adaptive Web Search Engines

M. Barouni-Ebrahimi
Faculty of Computer Science, University of New Brunswick, Canada

Web search engines efficiently surf the Internet and return the most relevant pages to the users' queries. However, the order of the recommended pages is not always in accordance with the users' priorities. The users needs to check the list of the recommended pages to find one of their interests. On the other hand, the queries sent by the users do not always corresponds to their intentions. The lack of user knowledge or unfamiliarity with the specific keywords and phrases in the domain knowledge leaves the user wondering about what phrases would be the most related ones to his desire. The contribution of our research is three folds. First, Complementary Phrase Recommender module suggests to the user a list of complementary phrases for his uncompleted query. Second, Related Phrase Advisor module provides a list of phrases related to the query segment that user has entered. These two modules guide the user to enter the more related phrases to his intention as a query. Third, Page Rank Revisor module refines the order of the recommended documents prepared by a conventional web search engine to help the user find the related web pages at top of the list.
This is joint work with Prof. Ali A. Ghorbani

Evaluating Web Search Quality

Maxim Gurevich
Department of Electrical Engineering, Technion, Haifa 32000, Israel.

Objectively assessing search quality is of great interest both to end-users and to search providers. Quality parameters like ranking quality, coverage of the web, index freshness, topic- and domain-specific coverage, and spam resilience are important for judging the effectiveness of search engines. Currently, search quality is evaluated mainly by manual techniques, using anecdotal test data. This makes the results difficult to reproduce and non-objective.
Random sampling is arguably the most efficient way to measure parameters on huge data sets, like search engines. Along sampling from the whole web/index of the search engine, sampling web pages from a given "topic" of the web/index (i.e., web pages in some domain, topic, language, etc.) may be interesting. Such samples can be used to measure the quality of search engines with respect to specific segments and domains. Evaluating ranking of search engines is another interesting area. Unlike current methods, which mainly rely on user studies, an automatically computable metric would be more statistically accurate and easy to reproduce. One idea is to use the latent human judgment in click-through

data. The metric may then be used to compare rankings of major search engines.

Use of k-cores to characterize graph local structure

John Healy
Dalhousie University

A k-core of a graph is the subgraph generated by recursively removing all nodes with degree $< k$. This can be thought of as a weaker version of a clique. k-cores are useful for pruning low importance vertices. ncreasing the value k eliminates nodes resulting in a component either remaining in our graph, splitting into multiple components, or being eliminated entirely. The resulting small directed acyclic graph, capturing the evolution of components as k increases, reveals the structure of a graph. We are using the k-core representation in order to find which of several available generative models is the best description of a real world graph, by developing a method of summarizing the component trees for statistical comparison.

The Chromatic Number of Complex Networks

Paul K Horn
University of California, San Diego

The chromatic number of a graph, denoted $\chi(G)$ is a graph invariant which is deeply tied to the structure of the graph, as well as other important graph properties such as the independence number and clique number. Here, we consider the chromatic number of complex networks by modeling complex networks as random graphs with given expected degree sequences through the $G(\mathbf{w})$ model introduced by Chung and Lu. We consider a graph with a more general degree distribution $\mathbf{w} = (w_1, \ldots, w_n)$; letting $w = (w_1 + \ldots + w_n)/n$ denote the average expected degree. We build upon work in a recent preprint of Frieze, Krivelevich and Smyth, giving in particular a condition guaranteeing that $\chi(G(\mathbf{w})) = \theta(w/\ln w)$; furthermore we show that if $\mathbf{w}$ fails this condition too badly, that indeed $\chi(G(\mathbf{w})) = \omega(w/\log w)$. We also give an improved lower bound on $\chi(G(\mathbf{w}))$.
We also investigate some related questions. Our conditions on when $\chi(G(\mathbf{w})) = \omega(w/\log w)$ seem to indicate a better indicator of the chromatic number. Still open is the question of finding an asymptotic value of $\chi(G(\mathbf{w}))$. Deeply related to this question are questions regarding the size and number of independent sets in $G(\mathbf{w})$. A deeper understanding of these issues is likely necessary to asymptotically determine $\chi(G(\mathbf{w}))$ and is also interesting in its own right.

Graph theory in interconnection networks

Navid Imani
Simon Fraser University

My main area of research lies somewhere between distributed computing and graph theory & enumerative combinatorics. I am interested in the problems arising in a wide variety of networks ranging from interconnection networks for multiprocessor systems and massively parallel systems to mobile and sensor networks and the WWW. My previous and on-going works address well-known problems in such networks such as Load Balancing (3 papers), Resource Placement (3 papers), Intrusion detection (4 Papers), Distributed Data-Clustering (3 papers), Studying Combinatorial Properties of Existing and Newly Proposed Networks (4 papers). Most of my papers involve in theoretical modeling of the above mentioned issues and thus my works have deep roots in graph theory, enumerative combinatorics and algorithms. Currently, I am involved in a research in network security where I study the probabilistic behavior of networks in the face of different types of failures using a combination of approaches from probability theory and combinatorics. Here, I consider different probabilistic fault patterns for nodes/links of the networks and calculate the effect of this fault on the overall network performance. The other problem I am envisioning is the probabilistic modeling of network disconnection assuming random node/link failures. I am hoping to propose similar solutions for the security problems of the web graph.

How NAGA uncoils: Searching with Relations and Entities.

Gjergji Kasneci
Max-Planck-Institut für Informatik, Saarbrücken, Germany

The everlasting enrichment of the Web with certain as well as uncertain and unstructured information calls for vertical search techniques which fulfill users' needs for querying the Web in a more precise way. Going one step beyond keyword search and allowing the specification of contextual concepts for keywords or relationships holding between them clears the way for new attractive and promising possibilities.

We present NAGA: A semantic search engine for the Web which exploits relationships between entities for precise query

specification and answering. NAGA's trump card is its ontological knowledge graph built on top of a refined data model which in turn serves as a basis for NAGA's query model and answer computation algorithms. NAGA extracts facts from Web pages and stores them into the above mentioned knowledge graph. Not only the extracted facts are recorded, but also a confidence measure for each fact is computed and maintained. NAGA provides a query language which can be capable of expressing queries ranging from simple keyword queries to complex graph queries which utilize regular expressions over relation names. NAGA's answer model is based on subgraph matching algorithms which in turn make use of intuitive scoring and ranking mechanisms. The approach we follow represents a general approach towards the semantic processing of information extracted from any unstructured text corpora.
This is joint work with Maya Ramanath, Fabian Suchaneck, and Gerhard Weikum.

Model of neighbourhoods in the web graph

Neil McKay
Dalhousie University

I first studied the web graph in the summer of 2005 as an undergraduate. As NSERC USRA summmer student I worked with Dr. David Pike examining a model of neighbourhoods in the web graph. We considered two nodes to be related if they each linked to each other. The goal was to achieve a finer macroscopic view of the web compared to the bow-tie structure of Broder, et al. (2000). This work resulted in a presentation given at the APICS Annual Math, Stats and Comp. Sci. conference. During the beginning of last summer I held a 12-week NSERC USRA again under the supervision of Dr. David Pike. I research the n-e.c. property in block-intersection graphs of balanced incomplete block designs (BIBDs). Most recently I completed a summer school course entitled Massive Networks and Internet Mathematics. We covered topics such as models for the web graph, various specific massive networks, G(n; p), the infinite random graph and algorithms such as Pagerank and HITS. The evaluation for the course was based on a project and presentation. My topic was finite examples of n-e.c. graphs.

Improving the Random-Surfer Model with Anonymized Traffic Data

Mark Meiss
Indiana University

Link-analytical algorithms for ranking Web search results such as Google's PageRank derive their power from the implicit statement of relevance made when the owner of one page decides to link to another. However, such methods are undermined by the fact that not all links are created equal: some are used much more often than others. The random-surfer model of PageRank assumes uniform distributions for starting locations, outgoing links to follow, and jump probabilities, but the behavior of actual surfers may be quite different. The aim of the present research is to gather large volumes of "click data" from anonymized packet captures of real HTTP sessions, analyze the extent to which this data does not reflect the random-surfer model, and develop a more sophisticated stochastic model in which the random distributions are based on the traffic patterns of actual users.

Neighborhoods in the Web Graph

Isheeta Nargis
Department of Computer Science, Memorial University of Newfoundland

The World Wide Web can be represented by a large directed graph in which each vertex corresponds to a web page, and in which there is an arc from one vertex to another if there is a hyperlink between the corresponding web pages. It is infeasible to store and manipulate the entire Web Graph, so we take a focused approach. Beginning with a specified web page, we determine which other pages are in close proximity to it, and then we construct the subgraph of the Web Graph that is induced by these pages (i.e., we construct a Neighborhood Graph for the given initial vertex). We investigate and report on certain properties of these neighborhood graphs.

Communities in Large Networks: Identification and Ranking

Martin Olsen
Department of Computer Science, University of Aarhus

We study the problem of identifying and ranking the members of a community in a very large network with link analysis only, given a set of a (few) representatives of the community.
We define the concept of a *community* justified by a formal analysis of a simple model of the evolution of a directed graph. We show that the problem of deciding whether non trivial communities exists is NP complete. Nevertheless experiments show that a very simple greedy approach can identify members of a community in the Danish part of the www graph

with time complexity only dependent on the size of the found community and its immediate surroundings.
We rank the members in a community by performing a computationally inexpensive calculation which is a "local" variant of the PageRank algorithm. The mathematical model behind the ranking is a small Markov Chain with the community as its state space forming a valuable basis for analyzing consequences of changes of the link structure.
Results are reported from a successful experiment on identifying and ranking Danish Computer Science sites.

Enhancing Data Analysis with Noise Removal

Gaurav Pandey
University of Minessota

Some very interesting recent studies have shown that protein interaction networks show very similar properties as the Web, such as the power law distribution followed by the degrees of the proteins in a network. The properties have also been utilized by studies that try to address important biological problems such as protein function prediction using techniques from other fields such as social network analysis and web search and mining. I believe that there is a rich potential for very useful research in this field of interaction network analysis via techniques from web search and mining, and an observation to this effect has also been made in my survey on protein function prediction. It is this problem that has motivated me to apply for the MoMiNIS Winter School. This school will provide me an opportunity to gain a deep insight into the field of modeling and mining of networked information spaces, such as the Web. I plan to leverage the techniques learnt from this school for the analysis of widely available protein interaction data during my thesis research. In many cases, these techniques would need significant modifications or enhancements, in order to match the characteristics of a biological network. Also, for several problems, these techniques may be combined with standard data mining approaches such as graph clustering algorithms. This rich suite of techniques will go a long way in the analysis of protein interaction networks, and the extraction of potential biological hypotheses that can be tested via wet-lab experiments.

Growing and classical protean graphs (new probabilistic models of the web)

Pawel Pralat
Department of Mathematics and Statistics, Dalhousie University

The web may be viewed as a graph each of whose vertices corresponds to a static HTML web page, and each of whose edges corresponds to a hyperlink from one web page to another. Recently there has been considerable interest in using random graphs to model complex real-world networks to gain an insight into their properties.
We propose an extended version of a new random model of the web graph in which the degree of a vertex depends on its age. We use the differential equation method to obtain basic results on the probability of edges being present. From this we are able to characterize the degree sequence of the model and study its behaviour near the connectivity threshold.
We present also the classical version of the model and characterize the limit distribution of the 'recovery time' for connectivity near the connectivity threshold, and the diameter of the giant component.
This is a joint work with Tomasz Luczak and Nicholas Wormald.

Probabilistic models for concept discovery in unstructured text data

Mahdi Shafiei
Faculty of Computer Science, Dalhousie University

Using probabilistic models for document and term clustering, document modeling and co-clustering has shown some major benefits over the traditional methods in recent years. In data mining research, these problems along with other problems including dimensionality reduction, topic segmentation, topic tracking and detection are closely related.These are also the fundamental building blocks of approaches to several applied problems including automatic summarization and machine translation. However, these problems have been approached independently of one another by the research community. I intend to bring all these interrelated problems under a single statistical model, and to exploit their interrelations. In my previous work, I have developed hierarchical Bayesian models for clustering terms and documents. By the topic segmentation capability embedded in the model, we hope to improve the clustering performance of the previous model on words and documents. Using the probabilistic Bayesian approach enables us to extend the proposed approach to a model capable of modeling topic tracking and shift in a principled way.

Shrack: A Pull-Only Peer-to-Peer Framework for Sharing and Tracking of Research Publications

Hathai Tanta-ngai
Faculty of Computer Science, Dalhousie University

We present Shrack—a pull-only peer-to-peer framework for document sharing and tracking. Shrack is designed to support researchers in forming direct collaborations to autonomously share and keep track of new research publications based on their interests. A pull-only information dissemination protocol is used for peers to learn about document metadata of new research publications from peers having similar interests. A user's interest is represented by an automatically learned profile. Document metadata are viewed as semi-structured documents. Peers first join the network using contacts acquired from real world collaboration, similar to exchanging email addresses or URLs. These contacts are used as initial peer neighbours. Each peer can use the disseminated information to build a local view of a semantic overlay network of peer interests, which represents groups of peers having similar semantic interests. Each peer can later use the semantic overlay network to find new contacts of peers having a particular interest, as well as search for documents archived by other peers. An overview architecture of the system and research challenges are presented.

Privacy in Databases

Dilys Thomas
Stanford University

The explosive progress in networking, storage, and processor technologies has resulted in an unprecedented volume of digital information. This has resulted in an increased real-time processing of this digital information in streaming systems. In concert with this dramatic and escalating increase in digital data and its real-time processing, concerns about privacy of personal information have emerged globally. The ease at which data can be collected automatically, stored in databases and queried effi- ciently over the internet has paradoxically worsened the privacy situation, and has raised numerous ethical and legal concerns [11, 12, 13, 28, 29, 30]. These concerns extend to the analytic tools applied to data. Problems arising from private data falling into malicious hands include identity theft, stalking on the web, spam etc. In the digital age, large amounts of confidential information are accessible to hackers or insiders. Safeguards to protect the privacy of individuals, and security of society are becoming crucial for the effective functioning of the internet. Privacy enforcement today is being handled primarily through legislation. We aim to provide technological solutions to achieve a tradeoff between data privacy and data utility.

Web Mining putting emphasis on Web Graph Evolution monitoring

Akrivi Vlachou
Athens University of Economics and Business

My research interests focus on web mining and in particular on link analysis and techniques for web graph representation. The web is a highly dynamic structure constantly changing. One of the biggest challenges is that of searching the vast amounts of web graph data. The research area of web search inherently involves the issue of page ranking. We address research problems related to the web graph evolution aiming at valid PageRank predictions and monitoring the web-graph change. Additionally, we envisage a compact representation of the web graph capitalizing on the changes of the web graph during time. Such a representation will be used to effectively answer historical queries.

Statistical Analysis of Dynamic Communication Graphs

Xiaomeng Wan
Dalhousie University

Communication networks can be modeled as a dynamic graph with time-varying edges. Real-life events cause communications that are unusual in either volume or pattern in the graph. Given such a dynamic graph with embeded events, can we detect when and where those events occur? The answer for this question is crucial for counter terrorism, network surveillance and traffic management. Most event detection methods only focus on network-wide events. However, events associated with only a few individuals are more common and of significant interest, as well. In this project, we explore a method to detect those events with only local impacts. We focus on three metrics to characterize communications from different viewpoints. Based on the variations of these metrics over time, we detect and characterize local events. Experiments on email data from our faculty show that these metrics are effective in identifying events, and the signals of the three metrics, combined in different ways, enables us to discriminate different types of events.

Dual Dynamic Programming and Reinforcement Learning

Tao Wang
University of Alberta

We investigate the dual approach to dynamic programming and reinforcement learning, based on maintaining an explicit representation of stationary distributions as opposed to value functions. A significant advantage of the dual approach is that it allows one to exploit well developed techniques for representing, approximating and estimating probability distributions, without running the risks associated with divergent value function estimation. A second advantage is that some distinct algorithms for the average reward and discounted reward case in the primal become unified under the dual. In this paper, we present a modified dual of the standard linear program that guarantees a globally normalized state visit distribution is obtained. With this reformulation, we then derive novel dual forms of dynamic programming, including policy evaluation, policy iteration and value iteration. Moreover, we derive dual formulations of temporal difference learning to obtain new forms of Sarsa and Q-learning. Finally, we scale these techniques up to large domains by introducing approximation, and develop new approximate off-policy learning algorithms that avoid the divergence problems associated with the primal approach. We show that the dual view yields a viable alternative to standard value function based techniques and opens new avenues for solving dynamic programming and reinforcement learning problems.

Sketching Landscapes of Page Farms

Bin Zhou
Simon Fraser University, Canada

The World Wide Web is a very large social network. It is interesting to analyze the general relations of web pages to their environment. For example, as rankings of pages have been well accepted as an important and reliable measure for the utility of web pages, we want to understand generally how web pages collect their ranking scores from their neighbor pages.
Such information is not only interesting but also important for a few web applications. (i) for Web spam, we can imagine identifying pages that receive a considerable amount of their ranking scores from bad pages; (ii) for Web page categorization, we could determine how much of page's ranking score comes from reputable pages from certain domains (e.g., the database and data mining community highly regards my page, but the network security community does not); (iii) for simple Web page characterization, it could be interesting to know that FedEx receives considerable link support from certain partner companies, etc.
In this project, we try to model the environment of web pages and analyze the general distribution of such environment. We study a novel web structure mining problem: mining page farms, and illustrate its application in link spamming detection. The general ideas and our major contributions so far are as follows.

- First, we study the page farm mining problem. A page farm is the set of pages contributing to (a major portion of) the PageRank score of a target page. We study the computational complexity of finding page farms, and show that it is NP-hard. Then, we develop a practically feasible greedy method to extract approximate page farms.
- Second, we empirically analyze statistics of landscapes of page farms using over 3 million web pages randomly sampled from the web. We have a few interesting findings.
- Third, we investigate the application of page farms in spamming detection. We propose two spamicity measures which can be used to detect spam pages. We evaluate our spamming detection methods using a real data set. The experimental results show that our methods are effective in detecting spamming pages.