707.000
Web Science and Web Technology
„Overview and Motivation“

Markus Strohmaier

Univ. Ass. / Assistant Professor
Knowledge Management Institute
Graz University of Technology, Austria

e-mail: markus.strohmaier@tugraz.at
web: http://www.kmi.tugraz.at/staff/markus
Web Science and Web Technology

- Welcome
- Motivation
- Introduction of Instructor and TA
- Course Modalities
- Background
In Spring 2008 ...

[Kaminsky] was looking at an error coded into the heart of the Internet's infrastructure.

This would allow him to reassign any Web address, reroute anyone's email, take over banking sites, or simply scramble the entire global system. The question was: Should he try it?

"Oh shit," he mumbled. "I just broke the Internet."
Click Fraud Index

Criminality

Now in its third year, the Click Fraud Index monitors and reports on data gathered from the Click Fraud Network. The Click Fraud Network provides statistically significant industry PPC data collected from online advertising campaigns for both large and small companies across all the leading search engines. Key findings from data reported for Q4 2008 include:

- The overall industry average click fraud rate grew to 17.1% for Q4 2008. That's up from 16.0% in Q3 2008 and from the 16.6% rate reported for Q4 2007.
- The average click fraud rate of PPC advertisements appearing on search engine content networks, including Google AdSense and the Yahoo Publisher Network, was 20.2%. That’s up from the 27.1% rate reported for Q3 2008 and down slightly from the 28.3% rate reported for Q4 2007.
- Traffic from botnets was responsible for 31.4% of all click fraud traffic in Q4 2008. That’s up from the 27.6% rate reported for Q3 2008 and the 22.0% rate reported for Q4 2007.
- In Q4 2008, the greatest percentage of click fraud originating from countries outside the U.S. came from Canada (7.4 percent), Germany (3.0 percent) and China (2.3 percent).

"Based on the data we tracked in Q4 2008, it seems that the online advertising industry is not immune to the growing tide of cybercrime during this recessionary period," said Tom Cuthbert, president of Click Forensics. "Both the overall click fraud rate and the rate of click fraud originating from botnets were the highest ever in Q4 2008. In addition, we've started to see old schemes like click farms reemerge. Advertisers should pay close attention to these types of threats in their online campaigns throughout the year."
Privacy in Social Networks [WWW 2009]

To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles

Elena Zheleva
Department of Computer Science
University of Maryland, College Park

Lise Getoor
Department of Computer Science
University of Maryland, College Park

Friendship network:

Social network groups:

- class labels (public profiles)
- unknown labels (private profiles)
Search (like it’s 1997!)

[http://web.archive.org/web/19981111183552/google.stanford.edu/]
Computers - another 10 years back (1987)

„Web science? Can you say that again?“
Motivation

“[…] As the Web has grown in complexity and the number and types of interactions that take place have ballooned, it remains the case that we know more about some complex natural phenomena (the obvious example is the human genome) than we do about this particular engineered one.”

A Framework for Web Science
Course team

• Instructor: Markus Strohmaier
• Teaching Assistants (planned):
  – Gabriele Zorn-Pauli
  – Christian Körner
  – Daniel Lamprecht
  – Ingo Holzmann

• e-mail addresses:
  – gabriele.zorn-pauli [at] student.tugraz.at
  – christian.koerner [at] tugraz.at
  – daniel.lamprecht [at] student.tugraz.at
  – ingo.holzmann [at] student.tugraz.at

• For general, course- or assignment-related questions, please use the newsgroup tu-graz.lv.web-science
About me

Education:
• 2002 - 2004 PhD. in Knowledge Management, Faculty of Computer Science, TU Graz
• 1997 - 2002 M.Sc., Telematik, TU Graz

Background:
• July 2007 - present: Ass. Prof. (Univ.Ass.), TU Graz, Austria
• 2006 - 2007 15 months Post-Doc, University of Toronto, Canada
• 2002 - 2006 Researcher, Know-Center, Austria
About me

**Research Background:**
- Web Research / Knowledge Management

**Research Interests:**
- Web Science with a focus on networks and Social Computing
- Intentional Structures and Representations on the Web

**Interesting topics for projects, Bachelor / Master thesis:**
- If you are interested in the topics of this course, it is likely that you are interested in doing a project / a thesis with me as well. Contact me to discuss opportunities.
Course Context

• 707.000 Web Science and Web Technology
  – 2nd year
  – Has been held before twice
  – Home assignments are different from last year’s course!

• Part of „Software Engineering & Business“
  – Bachelor studies, 6th semester
  – programming skills are required

• Your feedback is appreciated
Course Organization and Logistics

- **Lectures**
  Mondays 12:15 - 13:45,
  Feb 2009 - June 2009,
  Room HS i12 (Inffeldgasse 16b, Ground Floor)

- **Website:** [http://kmi.tugraz.at/staff/markus/courses/](http://kmi.tugraz.at/staff/markus/courses/)
- **Newsgroup:** tu-graz.lv.web-science
  - Please use the newsgroup for all questions related to the course

**Enroll!**
In order to obtain a grade, you need to enroll for this course until March 2 2009 via TUG online!

- **Weekly Readings**
  Password to access protected documents on the course website:
Grading

So how do you receive a grade in this course?

- 50% **home assignments**
  - Due dates for submissions are announced on the course website

- 50% **final exam**
  - On 23.6. 2008, no aids are allowed

- **Prerequisites:** Course „Einführung in die Strukturierte Programmierung“, **Java Programming Knowledge**

  In order to successfully complete the course, you need to have a score of 
  >= 51%

You can **cancel** your participation in this course anytime before the final exam: (will not result in a negative grade)
Grading

The following weights will be assigned to home assignments and the final exam (totalling 100%):

• **MatLab Exercises** (30%), *Individual Work*
• **Projects** (20%), *Team Work (3-4 students)*
  – **Option A**: Google Marketing Challenge (tight Timeline!, overlap with MatLab, fixed tasks, analytic)
  – **Option B**: Social Search Mashup (flexible Timeline, more creative tasks, implementation)
• **Final Exam**: 50%
  – „Nachklausur“
    precondition: at least 30% of points
Course Organization

To successfully complete the course, I recommend attending the weekly lectures, in which we

- **discuss papers** (read before class!)
- **go through the theory necessary for home assignments**
- **answer questions related to home assignments**
## Approximate Course Schedule

<table>
<thead>
<tr>
<th>Month</th>
<th>MatLab Exercises</th>
<th>Yahoo Boss Search Challenge</th>
<th>Google Online Marketing Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>March</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>April</td>
<td></td>
<td>Easter holidays</td>
<td>Overlap!</td>
</tr>
<tr>
<td>May</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>June</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Optional!
MatLab Exercises

• Tools
  • Matlab, ConExp, Pajek

• Networks
  • One-Mode Networks
  • Two-Mode Networks

• Analysis
  • Lattices
  • Metrics, such as Centrality, Redundancy
  • Ranking
  • Network simulators
  • Network infection

How many of you know MatLab?
In Brief:
• Build teams of 3-4 students
• Use At least 2 different APIs, Yahoo BOSS + 1 other (e.g. Twitter API)
• Build Your Own innovative Search Engine and

• Deploy it!
• Evaluate it until June 8th
• Write a brief report highlighting your idea (2 pages)

Winners:
• Determined by a panel of experts @ KMI
• Presentation of your search engine at the Web-Science course
• Linked from the course website

More details to follow, in the meantime also see: http://mashable.com/boss/
Alternative: Google Online Marketing Challenge

In Brief:
• Build teams of 3-4 students
• Find a partner with a website from industry
• Invest (real!) 200 US-Dollar in an:

Ad Campaign:
• Run it from
  May 1st – 21st 2009 (hard deadline)

Winners:
- global: Trip to Mountain View
- local: Trip to regional Google Office

More details to follow, in the meantime also see: http://www.google.com/onlinechallenge/
Approximate Course Schedule

<table>
<thead>
<tr>
<th></th>
<th>MatLab Exercises</th>
<th>Yahoo Boss Search Challenge</th>
<th>Google Online Marketing Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>March</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>April</td>
<td></td>
<td>Easter holidays</td>
<td>Overlap!</td>
</tr>
<tr>
<td>May</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>June</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Course Policies

• Class attendance and participation are voluntary, but highly recommended!
• Readings should be done before class
• All assignments are due at the beginning of the class on the due date
• Deadlines are sharp
• Assignment descriptions and lecture notes will be made available on the web
• Dishonesty (cheating, plagiarism)

For details see the course website:
http://kmi.tugraz.at/staff/markus/courses/707.000_web-science/
Course Topics

• World Wide Web
• What is network theory? Why is it relevant for the web?
• How do networks evolve?
• How do you search in networks?
• What are social parameters of networks?
• What are current web technologies?

But also e.g. a brief History of Smileys ;-) 
Simulations: e.g. [http://cmol.nbi.dk/javaapp.php](http://cmol.nbi.dk/javaapp.php)
How many of you know…

- 6 degrees of separation (small world problem)
- Power law networks
- Network generators
- The **Meaning** of PageRank? (E&A, Disk.Math.)
- Degree Distributions
- Galois Lattices
- …
Goals

• To equip you with the basic knowledge and tools for performing independent (network) analysis on the web
  – Knowledge:
    Network theory, search, mining, tagging, web technologies, applications, …
  – Tools:
    MatLab, Pajek, REST, simulators, network metrics
Preliminary Course Schedule

http://kmi.tugraz.at/staff/markus/courses/SS2009/707.000_web-science/
Non-Goals

In the research community, there is no broad consensus regarding the theoretical foundations of a „Science of the Web“ yet.

So therefore, the topics of this course are necessarily subjectively selective.

Instead of giving an authoritative account of web science, this course aims to give an overview of prominent, interesting and/or powerful research results generated by related fields so far.
Recommended Literature

There is no required text book for this course, however you might find it helpful to have a look at the following resources:

- **Six Degrees - The Science of a Connected Age**, Duncan J. Watts, 2004
- **Web Dragons**, Ian Witten et al, 2007
- **Social Network Analysis - Methods and Applications**, Stanley Wasserman and Katherine Faust, 1995
- **Graph Theory**, Reinhard Diestel, Electronic Edition, 2005 ([free PDF download](#))
Questions?

Raise them **NOW!**

Or ask them later:
  - At the end of each class
  - Via newsgroup

*(now would be a good time though!)*
Let’s start!
- Science and the Web -
The Web: Looking Back, Looking Forward
Tim Berners-Lee

Talking about Web Science
(~ 70mins)

The Web: Looking Back, Looking Forward
[Berners-Lee 2007]

The process of web science

Science
Engineering

Values

Creativity

Issues

Idea

Design

Social

Tech.

Macro

Micro

Analyze

Complexity

6,000,000,000 people

10,000,000,000 web pages

Berners-Lee
Lovelace lecture 2007-03-13

Markus Strohmaier
2009
The Web Today

Web 2.0 ... The Machine is Us/ing Us

http://www.youtube.com/watch?v=6gmP4nk0EOE

How do the topics addressed in this movie relate to a Science of the Web?
Motivation

A reported number of 900 Mio people (that is roughly one out of seven people on earth!) watched a video of a previously unknown, video amateur, teenage starwars fan.

How is this possible? How does information spread on the web? What are the effects on individuals and society?

http://entertainment.timesonline.co.uk/tol/arts_and_entertainment/tv_and_radio/article650932.ece
A Brief Overview of the Web
[Berners Lee et al 1994]

- Vision: the W3 operates without regard to
  - Where information is stored
  - How information is stored or
  - What system is used to manage it
- Documents referring to each other by links
- Analogy to spiders’ construction: the web
- Hypertext paradigm
  - Sensitive parts of text representing links
  - A link is followed by mere pointing and clicking (or typing a ref. Nr.)
  - No primary focus on search
- Hypertext links may be made to any data in non-W3 servers (FTP, Gopher, WAIS or internet news) as W3 clients have the ability to present all such data as hypertext.

- The World Wide Web combines Hypertext and Search

the web != internet
The web: Presentation and Extraction
[Berners Lee et al 1994]

The architecture of W3 (fig. 2) is one of browsers (clients) which know how to present data but not what its origin is, and servers which know how to extract data but are ignorant of how they will be presented. Servers and clients are unaware of the details of each other’s operating system quirks and exotic data formats.

All the data in the Web is presented with a uniform human interface (Fig. 3). The documents are stored (or generated by algorithms) throughout the internet by computers with different operating systems and data formats. Following a link from the SLAC home page (the entry into the Web of a SLAC user) to the NIKHEF telephone book is as easy and quick as following the link to a SLAC Working Note.

Fig. 2: Architecture of W3
The web
[Berners Lee et al 1994]

Fig 1. The basic hypertext model is enhanced by searches.
Features of the web
[Berners Lee et al 1992]

Features to note are:-

- Information need only be represented once, as a reference may be made instead of making a copy;

- Links allow the topology of the information to evolve, so modeling the state of human knowledge at any time without constraint;

- The web stretches seamlessly from small personal notes on the local workstation to large databases on other continents;

- Indexes are documents, and so may themselves be found by searches, and/or following links. An index is represented to the user by a “cover page” which describes the data indexed and the properties of the search engine.

- The documents in the web do not have to exist as files: they can be “virtual” documents generated by a server in response to a query or document name. They can therefore represent views of databases, or snapshots of changing data (such as the weather forecast, financial information, etc).
Historical Vision of the Web

Is a space in which

- **Resources** are identified by Uniform Resource Identifiers (URIs)
- **Protocols** support interaction between agents (HTTP)
- **Formats** represent information resources (HTML)
Uniform Resource Identifier

- Resources may be anything that can be linked to or spoken of
  - Resoures can contain a reference to another resource
- Identifiable, but not necessarily retrievable
  - (e.g. protected access)
- A single global system of identifiers
  - Each URI ideally identifies a single resource in a context-independent manner
- URIs act as names and addresses
- URIs require institutions
  - E.g. the registry that handles domain names
HTTP & HTML: High Level Overview
http://www.w3.org/Protocols/HTTP/HTTP2.html

HTTP: A protocol that is basically stateless, a transaction consisting of
• Connection
  – The establishment of a connection by the client to the server - when using
    TCP/IP port 80 is the well-known port, but other non-reserved ports may be
    specified in the URL;
• Request
  – The sending, by the client, of a request message to the server;
• Response
  – The sending, by the server, of a response to the client;
• Close
  – The closing of the connection by either both parties.

HTML: A representation format
• Idea: Decoupling of content and representation
• Cues for graphical presentation of content
Why Web Science?

- Dynamics and evolution
- The “deep web” (resources not available by robots)
- Sampling, lack of complete enumeration
- Scale (e.g. “What’s the percentage of web pages updated daily?”)
- Search (e.g. “What’s the percentage of web pages indexed by search engines?)
- Web topology
- Artifacts of social interaction (weblogs, etc), web sociology
- …
Science (in a nutshell)

What type of question are you asking?

→ **Existence:**
  - Does X exist?

→ **Description & Classification**
  - What is X like?
  - What are its properties?
  - How can it be categorized?
  - How can we measure it?
  - What are its components?

→ **Descriptive-Process**
  - How does X work?
  - What is the process by which X happens?
  - In what are the steps as X evolves?
  - How does X achieve its purpose?

→ **Descriptive-Comparative**
  - How does X differ from Y?

→ **Relationship**
  - Are X and Y related?
  - Do occurrences of X correlated with occurrences of Y?

→ **Causality**
  - Does X cause Y?
  - Does X prevent Y?
  - What causes X?
  - What effect does X have on Y?

→ **Causality-Comparative**
  - Does X cause more Y than does Z?
  - Is X better at preventing Y than is Z?
  - Does X cause more Y than does Z under one condition but not others?

→ **Design**
  - What is an effective way to achieve X?
  - How can we improve X?
What could theories for the web look like?

Some Simple Examples:

• Every page on the web can be reached by following less than 10 links. (True/False/Depends?)
• 1%-4% of users express their search queries in the form of goals such as “increase adsense revenue” (True/False/Depends?)
• The average number of words per search query is more than 3 (True/False/Depends?)
• A wikipedia page contains, on average, 0.03 false facts (True/False/Depends?)

Can these statements be easily validated? Are these good theories? What constitutes good theories?
Some Quality Characteristics of Theories

- Clarity
- Simplicity
- Predictive Power
- Explanative Power
- Utility
- Testability
- Falsifiability (vs. Falsification)
A significant part of this course will focus on network theory.

- Graph theory vs. Network theory
  - While graph theory focuses on mathematics, network theory focuses on networks that can be observed in the „real world“
  - Evolution of networks

- There are many different forms of networks available on the net

- Can you name a few of them?
The Web as a Network of Related Sites

http://www.touchgraph.com/TGGoogleBrowser.html
(based on Google’s „related sites“ functionality)
The Web as a Network of Search Results

http://www.kartoo.com (search for „web2.0“)
Delicious as a Network of tags

![Image of a network diagram with tags like "freetime", "webdesign", and "travel" connected by lines]

**Fig. 1.** The del.icio.us tags associated through co-occurrence on items and the clusters emerging

| Table 1. The five main clusters of interest based on the Concept-Object network |
|-----------------------------|---------------------------------------------------------------|
| **travel**                  | cote, provence, villa, azur, mas, holiday, vacation, tourism, france, heritage |
| **business**                | venture, capital, enterprise, up, start, venture, newspaper, capital, Segev, pitango, vc |
| **free time**               | procrastination, info, advice, gtd, life, notes, planning, daily, reading, forums |
| **sex**                     | hot, to, street, pictures, on, photos, free, celeb, adult, lesbian |
| **web design**              | design, designer, webdesign, premium, logo, logos, dreamweaver, templates, best, good |
The Blogosphere as a Network of Blog Posts

In a model (framework) for weblog research it was suggested that one should look at five dimensions to study weblogs. This post shows that one can obtain a fascinating peek into the blogosphere by looking at just two dimensions: [links, persons]. Perhaps it is an idea to also add time so that we can see whether the yellow and pink posts occur before (this is possible), during or after the conversation.
User Goals

- Fall in love
- Make new friends
- Be a better friend
- Be a good friend
- Get a tattoo
- Be debt free
- Buying a house
- Learn Japanese
- Don't eat fast food
- Live a healthy lifestyle
- Lose weight
- Lose 10 pounds
- Lose 20 pounds
- Get in shape
- Exercise 3x a week
- Own a house
- Learn Chinese
- Learn more Japanese
- Travel to every continent
- Travel the world
- Write a book
- Write a good novel
- Publish a short story
- Finish my novel
- Have children
- Enjoy life
- Identify 100 things that make me happy
Web Science
www.webscience.org
Some Course Highlights
An Experimental Study of the Small World Problem [Travers and Milgram 1969]

A Social Network Experiment tailored towards
• Demonstrating
• Defining
• And measuring
Inter-connectedness in a large society (USA)

A test of the modern idea of “six degrees of separation”
Which states that: every person on earth is connected
to any other person through a chain of acquaintances
not longer than 6
Some Course Highlights
Some Course Highlights

Folksonomy Analysis
Check

• Is there anything else you want to know w.r.t. this course?
Any further questions?

Have a good start in the new semester!
- See you next week