

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: <u>markus.strohmaier@tugraz.at</u> web: <u>http://www.kmi.tugraz.at/staff/markus</u>



Web Science and Web Technology

- Welcome
- Motivation
- Introduction of Instructor and Teaching Assistants
- How this course is organized (and how you obtain a grade)
- Introduction to the course
- Some course highlights

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." TECH BIZ : PEOPLE 题

Secret Geek A-Team Hacks Back, Defends Worldwide Web

By Joshua Davis 🖂 👘 11.24.08

WIRED MAGAZINE: 16.12

"The first thing I want to say to you," Vixie told Kaminsky, trying to contain the flood of feeling, "is **never, ever repeat what you just told me over a cell phone**."

Kaminsky was alone in his Seattle apartment when he discovered a security vulnerability that could leave banks, online retailers, and ISPs open to hackers. *Photo: John Keatley*

From that moment on, they would talk only on landlines, in

person, or via heavily encrypted email. Secrecy was critical.

They had to find a solution before the problem became public.

http://www.wired.com/techbiz/people/magazine/16-12/ff_kaminsky http://www.youtube.com/watch?v=R-SSVxsH7vw 3

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 28.2%. 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 vear."

How can we identify spam, link farms and click fraud on the web?

Markus Strohmaier



Random Social Connections

How do random social graphs differ from "real" social networks?



http://vimeo.com/9669721

http://bits.blogs.nytimes.com/2010/02/13/chatroulettes-founder-17-introduces-himself/



Privacy

To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles [WWW 2009]





6 degrees of separation? http://arxiv.org/PS_cache/arxiv/pdf/0803/0803.0939v1.pdf

- 30 billion conversations among 240 million people of Microsoft Messenger
- Communication graph with 180 million nodes and 1.3 billion undirected edges
- Largest social network constructed and analyzed to date (2008)



Markus Strohmaier

Figure 14: (a) Communication among countries with at least 10 million conversations in June 2006. (b) Countries by average length of the conversation. Edge widths correspond to logarithms of intersity of links.

The Web in 2007





Total Sites Across All Domains August 1995 - September 2007

(courtesy, www.opte.org)

Sept 2007, Netcraft



Search (like it's 1997!)

[http://web.archive.org/web/19981111183552/google.stanford.edu/]

Google!		
Search the web using Google! 10 results Google Search I'm feeling lucky Index contains ~25 million pages (soon to be much bigger)		
<u>About Google!</u> <u>Stanford Search Linux Search</u>		
Get Google! updates monthly!		
your e-mail Subscribe Archive		
Copyright ©1997-8 Stanford University		



Computers - another 10 years back (1987)



**** COMMODORE 64 BASIC V2 **** 64K RAM SYSTEM 38911 BASIC BYTES FREE <u>R</u>EADY.





"Web science? Can you say that again?"



2010



2010



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

T. Berners-Lee and W. Hall and J. A. Hendler and K. O'Hara and N. Shadbolt and

D. J. Weitzner Foundations and Trends® in Web Science 1 (2006)



Course team

- Instructor: Markus Strohmaier
- Teaching Assistants:
 - Ingo Holzmann
 - Florian Klien
 - Christian Körner
 - Jan Pöschko
- e-mail addresses:
 - klien at@ student.tugraz.at
 - christian.koerner at@ tugraz.at
 - poeschko 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
 - 3rd year
 - Has been held before thrice
 - 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, Mar 2010 - June 2010, Room HS i12 (Inffeldgasse 16b, Ground Floor)

- Website: <u>http://kmi.tugraz.at/staff/markus/courses/</u>
- 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 15 2010 via TUG online!

• Weekly Readings

Password to access protected documents on the course website:





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 28.6. 2010, no aids are allowed

- **Prerequisites**: Course "Einführung in die Strukturierte Programmierung", **General 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:

just don't show up at the final exam, this will **not** result in a negative grade



Grading

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

- MatLab/Octave Exercises (30%), Individual Work
- **Projects** (20%), *Team Work* (3-4 students)
 - Details to be announced during this course
- **Final Exam**: 50%



Policies

- **Course documents**: Assignment descriptions and lecture notes will be made available on the course website
- **Deadlines**: Late submissions (same day) will result in a loss of 1/3 of all your points for this assignment. After the day of the deadline, no points will be awarded.
- **Plagiarism**: By submitting home assignments, you agree that your work will be processed by plagiarism tools.
- Nachklausur: Participating in the Final Exam is a prerequisite for every student who wants to obtain a positive grade for this course. Because you can already obtain 50% of the points during Home Assignments (>=51% are necessary overall), no general "Nachklausur" will be offered. (Exception: illness, see details on course website)

For all course policy details, please see the course website: http://kmi.tugraz.at/staff/markus/courses/SS2010/707.000_web-science/

Markus Strohmaier



Plagiarism

You are **allowed** to discuss home assignments with colleagues, but You are **not allowed** to jointly work on the assignments, copy solutions or share code.

We will apply sophisticated plagiarism detection software to your submissions.



Consequence: students could not complete the course in this year (repeat next year).



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



Course 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





Home Assignments HA1.1-1.5

- Tools
 - Matlab/Octave, ConExp, Pajek
- Networks
 - One-Mode Networks
 - Two-Mode Networks
- Analysis
 - Lattices
 - Metrics, such as Centrality, Redundancy
 - Ranking
 - Network simulators
 - Network infection



Home Assignment 1.1 will be handed out in week 3!



Home Assignments HA2

- To be announced and released Mid of May
- Team work (3-4 students)



Approximate Course Schedule

	MatLab/Octave Exercises	Project Assignments	
March	Ongoing		
	submission of home assign.!		
April	Easter	holidays	
May			
June			



Home Assignments

- Have been somewhat revised from last year's course
 -> some are completely new
- They might need clarification or refinement once they are handed out.
- Changes and/or updates to the assignments will occur
- Once they occur, they will be announced in the newsgroup, it is your responsibility to obtain them!



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. <u>http://cmol.nbi.dk/javaapp.php</u>



How many of you know...

- 6 degrees of separation (small world problem)
- Power law networks
- Network generators
- The Meaning of PageRank?
- Degree Distributions
- Galois Lattices
- ...



Preliminary Course Schedule

http://kmi.tugraz.at/staff/markus/courses/SS2010/707.000_web-science/



Non-Goals

In the research community, there is **no 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 authorative 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 resource:

- <u>Networks, Crowds, and Markets</u>: Reasoning About a Highly Connected World, by David Easley and Jon Kleinberg, 2010 (<u>free online book</u>)
- Also see the resources listed on <u>http://kmi.tugraz.at/staff/markus/courses/SS2010/707.000_web-science/</u>



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)

<u>http://www.w3.org/2007/Talks/0313-bcs-tbl/</u> (slides)
 <u>http://www.bcs.org/server.php?show=nav.9996</u> (lecture)
 <u>http://video.google.com/videoplay?docid=533364838415066499</u>
 <u>2&hl=en-GB</u> (lecture)



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







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

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



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 <u>uniform human interface</u> (Fig. 3). The documents are stored (or generated byalgorithms) 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 <u>represented once</u> , as a reference may be made instead of making a copy;	Uniqueness
	Links allow the topology of the information to evolve, so modeling the state of human knowledge at any time without constraint;	Networkability
	The web stretches seamlessly from small personal notes on the local workstation to large databases on other continents;	Scalability
	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.	Indexability
•	The documents in the web <u>do not have to exist as files</u> : 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).	Adaptability / Customizability



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 contextindependent 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 div class="banner">div class="banner">
- Cues for graphical presentation of content

<div>

<map name="introLinks" id="intro <div class="banner"> <a class Visitors | <a class="banner! <a class="bannerLink" title="Cor



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 could theories for the web look like?

A few examples of assertions:

- Every page on the web can be reached by following less than 10 links. (True/False/Depends?)
- A wikipedia page contains, on average, 0.03 false facts (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?)

Can these statements be easily validated? Can they lead to good/interesting theories? What constitutes good theories?



Some Quality Characteristics of Theories

- Clarity
- Simplicity
- Predictive Power
- Explanative Power
- Utility
- Testability
- Falsifiability (vs. Falsification)



Networks

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

ni nanozecine freetime googe texonomy to texonomy seo ka	Table 1. The five main clusters of in- terest based on the Concept-Object net-		
tips procrastination solagging photo advertising picasa firefox	work		
is webdesign webdesign bah review obiethreewer script operations de schore bool filter and the schore bool of the schore bool	\mathbf{travel}	cote, provence, villa, azur	
to webmaster pp determine powerpoint isoti duronakophan used in a substance privice film privice film inst website privice film inst used inst us		tourism, france, heritage	
tick quy a set Street	business	venture_capital, enterprise	
wall professional ted of the transmission of t		paper, capital, Segev, pi-	
300 1 8 motive and the A van with up we trans identify we take the average of the	free time	tango, vc procrastination, info, ad-	
sub woltegerin (head so war about bin travel bin Rand Belanital ma soart site and so the soart state of the soart		vice, gtd, life, notes, plan- ning, daily, reading, forums	
usa di anti anti interna di an	sex	hot, to, street, pictures, on photos. free. celeb. adult	
nyt		lesbian	
	web design	design, designer, webde-	
Fig. 1. The del.icio.us tags associated		sign, premium, logo, logos,	
through co-occurrence on items and the		dreamweaver, templates	
clusters emerging		best, good	

Markus Strohmaier



The Blogosphere as a Network of Blog Posts



Courtesy of http://anjo.blogs.com/





Web Science www.webscience.org



Markus Strohmaier

2010



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



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? How can we model this? What are the effects on individuals and society?

http://entertainment.timesonline.co.uk/tol/arts_and_entertainment/tv_and_radio/article650932.ece







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