Social Web and Bibliometrics

University of Koblenz-Landau,
SS 2013

York Sure-Vetter
# Preliminary Timetable

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Datum</th>
<th>Uhrzeit</th>
<th>Raum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tu</td>
<td>16.04.</td>
<td>16:00-18:00</td>
</tr>
<tr>
<td>2</td>
<td>Tu</td>
<td>30.04.</td>
<td>16:00-18:00</td>
</tr>
<tr>
<td>3</td>
<td>Tu</td>
<td>14.05.</td>
<td>16:00-18:00</td>
</tr>
<tr>
<td>6</td>
<td>Tu</td>
<td>04.06.</td>
<td>16:00-18:00</td>
</tr>
<tr>
<td>7</td>
<td>We</td>
<td>05.06.</td>
<td>14:00-16:00</td>
</tr>
<tr>
<td>8</td>
<td>Tu</td>
<td>11.06.</td>
<td>16:00-18:00</td>
</tr>
<tr>
<td>9</td>
<td>We</td>
<td>12.06.</td>
<td>14:00-16:00</td>
</tr>
<tr>
<td>10</td>
<td>Tu</td>
<td>25.06.</td>
<td>16:00-18:00</td>
</tr>
<tr>
<td>11</td>
<td>Tu</td>
<td>02.07.</td>
<td>16:00-18:00</td>
</tr>
<tr>
<td>12</td>
<td>Tu</td>
<td>09.07.</td>
<td>16:00-18:00</td>
</tr>
<tr>
<td>13</td>
<td>We</td>
<td>10.07.</td>
<td>14:00-16:00</td>
</tr>
<tr>
<td>Buffer</td>
<td>Tu</td>
<td>16.07.</td>
<td>16:00-18:00</td>
</tr>
<tr>
<td>Buffer</td>
<td>We</td>
<td>17.07.</td>
<td>14:00-16:00</td>
</tr>
</tbody>
</table>
The Web in 2007

The diagram shows the total number of sites across all domains from August 1995 to September 2007. The graph indicates a significant increase in the number of sites, growing from 136 million in September 2007 to 55 million in the same period. The data is courtesy of Netcraft.
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?“
Activity of #streams on Twitter over time

- Jan 9 Earthquake in California
- Earthquake in Haiti
- Aftermath of the earthquake in Haiti

Koblenz, Summer 2013
How did you wake up today?

Twitter confirms it:
people tend to wake up in a good mood
and are happiest on weekends.
(Cornell University, N: 2.4 million people in 84 countries)
Who is ED?

Franklin D. Roosevelt: Fireside chats
John F. Kennedy: Television
Barack Obama: Internet-video, Twitter and Facebook
What is gairfam?

Global Analysis of Intimate Relationships and FAMily dynamics (See also: Pairfam project … ;-) …)

Info: Gairfam data collection started in 2011

http://www.facebook.com/about/timeline
Random Social Connections

How do random social graphs differ from „real“ social networks?

http://vimeo.com/9669721

Privacy

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

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)
6 degrees of separation?

- 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)
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
About me (York)

I am professor for applied computer science at the University of Koblenz-Landau (in Koblenz) and co-head the Institute WeST – Web Science and Technologies.

However, my main duty is being President of the GESIS – Leibniz Institute for Social Sciences.

I have previously e.g. worked for SAP AG and visited Stanford University.
You may also ask …

What is GESIS doing?

How is this related to the University of Koblenz-Landau?

How comes you have two jobs?

*We’ll jointly work on the question: Why am I here? ;-)"
Course Topics

• *Some* World Wide Web
• What is network theory? Why is it relevant for the web?
• How do networks evolve?
• What are social parameters of networks?
• What are current (social) web technologies?
• How do researchers in particular collaborate?
• How do they publish and how can you measure joint research activities?
Aim of the lecture

Understanding of the typical regularities of the social web

Properties of the social web

Understanding of bibliometrics and its regularities

Similarity of social web and bibliometric regularities
How many of you know…

• 6 degrees of separation (small world problem)
• Power law networks
• Network generators
• The Meaning of PageRank?
• Degree Distributions
• …
Non-Goals

In the research community, there is no consensus regarding the theoretical foundations of a „Science of the Web“ yet (a bit more consensus is achieved wrt. „Social Web“, though).

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

Instead of giving an authoritative account of Web Science and Social Web, 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:


- Also see the resources listed on [http://kmi.tugraz.at/staff/markus/courses/SS2010/707.000_web-science/](http://kmi.tugraz.at/staff/markus/courses/SS2010/707.000_web-science/)
Acknowledgements

A BIG THANKS to Markus Strohmaier from the Knowledge Management Institute of the TU Graz!

In particular, he is frequently giving a full lecture on „Web Science and Web Technology“. Many slides are based on his lecture.

Please find out more about Markus at: http://www.kmi.tugraz.at/staff/markus
Questions?

Raise them NOW!

Or ask them later:
• During or at the end of each class

(now would be a good time though!)
Let’s start!
- The Web and The Social Web –
Background

„Facebook, SchülerVZ und Co. sind nicht nur die Lieblings-Treffpunkte von Abermillionen Surfern. Die sozialen Netzwerkdienste sind im Begriff, sich als die wichtigsten Torwächter für den Zugang zum Internet generell zu etablieren. “c‘t 7/10

FB has 800+ Mio. active users

Users of social networks spend more time online than online searchers (more then 15 pages views)
Background

„The term "Web 2.0" (2004–present) is commonly associated with web applications that facilitate interactive information sharing, interoperability, user-centered design, and collaboration on the World Wide Web. Examples of Web 2.0 include web-based communities, hosted services, web applications, social-networking sites, video-sharing sites, wikis, blogs, mashups, and folksonomies.

A Web 2.0 site allows its users to interact with other users or to change website content, in contrast to non-interactive websites where users are limited to the passive viewing of information…” Wikipedia
The Web

Some facts:

1989  Web Idea
1994  Amazon founded
1998  Google founded
2001  .Com Bubble
2004  Facebook founded
2008  Obama becomes president due to Web strategy
2009  SAP defines Web strategy

1. Generation: Advertising & Buy/sell
2. Generation: Live and Work
The Web: Looking Back, Looking Forward
Tim Berners-Lee

Talking about Web Science
(~ 70mins)

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

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 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)
URI

Uniform Resource Identifier

- Resources may be anything that can be linked to or spoken of
  - Resources 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?

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 $\leftarrow$ HOMEWORK!)
The Social Web is … Networks!

Some 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

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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>travel</td>
<td>cote, provence, villa, azur, mas, holiday, vacation, tourism, france, heritage</td>
</tr>
<tr>
<td>business</td>
<td>venture_capital, enterprise, up, start, venture, newspaper, capital, Segev, pitango, vc</td>
</tr>
<tr>
<td>free time</td>
<td>procrastination, info, advice, gtd, life, notes, planning, daily, reading, forums</td>
</tr>
<tr>
<td>sex</td>
<td>hot, to, street, pictures, on, photos, free, celeb, adult, lesbian</td>
</tr>
<tr>
<td>web design</td>
<td>design, designer, webdesign, premium, logo, logos, dreamweaver, templates, best, good</td>
</tr>
</tbody>
</table>
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.

Courtesy of http://anjo.blogs.com/
Web Science
www.webscience.org
The Social Web

Trend: more & more users spend a significant amount of time in social networks.

Koblenz, Summer 2013
The Social Web

More and more types of data are available in social networks

Koblenz, Summer 2013
The Social Web

Professional contacts and personal updates.
The Social Web

Neue Mitglieder

- Margret Schneeberger
  Akzo Nobel Deco GmbH

- Andreas Bronner
  Stiftung St. Franziskus Heiligenkron

- Thomas Schwebel
  BASF IT Services

- Peer Moritz
  Ingenieursgesellschaft P.A.T.H. mbH

Mitglieder, die Sie kennen könnten

- Hansgeorg Langhorst
  HISolutions AG

- Constanze Beierlein
  GESIS - Leibniz-Institut für Sozialwissenschaften

- Holger Pahl
  ZUMA

- Tom Kunze
  HISolutions AG

Besucher meines Profils

- Nur für Premium-Mitglieder
  Jetzt upgraden

- Nur für Premium-Mitglieder
  Jetzt upgraden

- Nur für Premium-Mitglieder
  Jetzt upgraden

- Nur für Premium-Mitglieder
  Jetzt upgraden

Kontakte meiner Kontakte

- Marc Methfessel
  The Linde Group

- Katrin Pehle
  Universal Pictures Germany

- Janette Klauck
  Preventas Hamburg

- Sandra Gruber
  Oerd Bär GmbH, Heilbronn

Professional contacts and personal updates:

- Contacts from contacts
- Visitors of the profile
- Applications
- Jobs

Jobs, die zu Ihnen passen könnten

- Medizinprodukte Berater/Homo Care
  Personalberatung Wagner

- Berater/Business Analyst (w/m)
  Finsyskoplan AG

aktuell | Meine Kontakte | Letzte Updates

Elastien Gogolin fragt

Welche weiteren Ausbildungsmöglichkeiten...
The Social Web

Home

ExLibrisGroup Congratulations to Keio University in Japan on going live with Aleph and KOSMOS, powered by Primo, http://tinyurl.com/zbuzt4c
22 minutes ago via web

about 2 hours ago via web

mstephens7 I favorited a YouTube video -- Monitoring NLM.mp4 http://youtu.be/lRufHj5xPMa
about 3 hours ago via Google

danbri http://wwwics.uci.edu/~lopes/dv/dv.html is just what I was looking for! http://wiki.foaf-project.org/WikiBr/ChirpChirp :) about 4 hours ago via identica


ivan_herman Back from a long series of meetings at MIT. Two tiring but enriching weeks. . .
about 5 hours ago via Power Twitter

Twitter-Blog n, a nifty place to read official news from Twitter.

News in 140 characters ...
Properties of the Social Web

Typical Social Web data
Profile details: age, location, sex, other data
Connectivity: friends, groups, memberships
Content: items, tagging, comments, edits, ratings, statistics
Technical: public API, embedding, RSS, messages
Typology

Categorization of Services (Thelwall, 2009):

1. **Socialising** … are designed for recreational social communication between members.

2. **Networking** … primarily designed for non-social interpersonal communication.

3. **Social navigation** … have social network features but use them primarily as a way to help users find a particular type of information or resource.
Typology

(Thelwall, 2009)
## Social Web Analysis

(Thelwall, 2009)

<table>
<thead>
<tr>
<th>Alexa global</th>
<th>Global Alexa rank*</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouTube</td>
<td>3</td>
<td>Video sharing SNS</td>
</tr>
<tr>
<td>MySpace</td>
<td>6</td>
<td>Socialising SNS</td>
</tr>
<tr>
<td>Facebook</td>
<td>8</td>
<td>Socialising SNS</td>
</tr>
<tr>
<td>Orkut</td>
<td>11</td>
<td>Socialising SNS (Google)</td>
</tr>
<tr>
<td>Hi5</td>
<td>19</td>
<td>Socialising SNS</td>
</tr>
<tr>
<td>V Kontakte</td>
<td>30</td>
<td>Russian socialising SNS</td>
</tr>
<tr>
<td>Flickr</td>
<td>39</td>
<td>Image sharing navigational SNS</td>
</tr>
<tr>
<td>Friendster</td>
<td>40</td>
<td>Socialising SNS</td>
</tr>
</tbody>
</table>

Social network sites in the top 100 Internet sites, according to Alexa (May, 2008).
Social Web Analysis: Who?

Ages of MySpace members in the two data sets, as tested in March 2008.

(from Thelwall, 2009)
Social Web Analysis: What for?

Declared purpose for using MySpace, broken down by gender.

(from Thelwall, 2009)
Finally … Some Highlights
Some Social Web 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 Social Web Highlights

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

Koblenz, Summer 2013
Some Social Web Highlights

Folksonomy Analysis

- London: Terrorism, News, UK, Bombs, Politics, Terrorism, Explosions
- Travel: Music, Current Affairs, Food and Drink, Photography, Books, Weblogs
- France: Europe, Paris, Germany, European Union, Photo, Travel, Politics, French
- Photos: Family, Travel, Babes, Flickr, Photography, Personal Life, Music, News

Koblenz, Summer 2013
Any further questions?

Have a good start!
... and enjoy the beginning of the summer in Koblenz!