Computational Social Science

Course #04199, module 04IN2042

Markus Strohmaier

Prof. Dr.

WeST / IWVI U. of Koblenz-Landau, Germany

e-mail: markus.strohmaier@gesis.org

web: http://www.markusstrohmaier.info

Computational Social Science

- Welcome

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

Computational Social Science

… Social Science?

What is

… Computer Science?

R. P. Feynman on social sciences

(1981)
http://www.youtube.com/watch?v=IaO69CF5mbY&list=FLg5a1LWfGRn4grwWIwCzcLA

Richard Feynman on Computer Science — Talk at Bell Labs (19…

(1985)
http://www.youtube.com/watch?v=IL4wg6ZAFIM&list=FLg5a1LWfGRn4grwWIwCzcLA

Social Science: a simplification

Social Scientists are interested in how people

- think/feel/behave in social situations (social psychology),
- relate to each other (sociology),
- govern themselves (political science),
- handle wealth (socio-economics) and
- create culture (anthropology).

Social and Computer Science in the 21st century

New kinds of data (macro-scale)

Human mobility in societies

Check-ins (Foursqaure, Gowalla, Twitter, …)

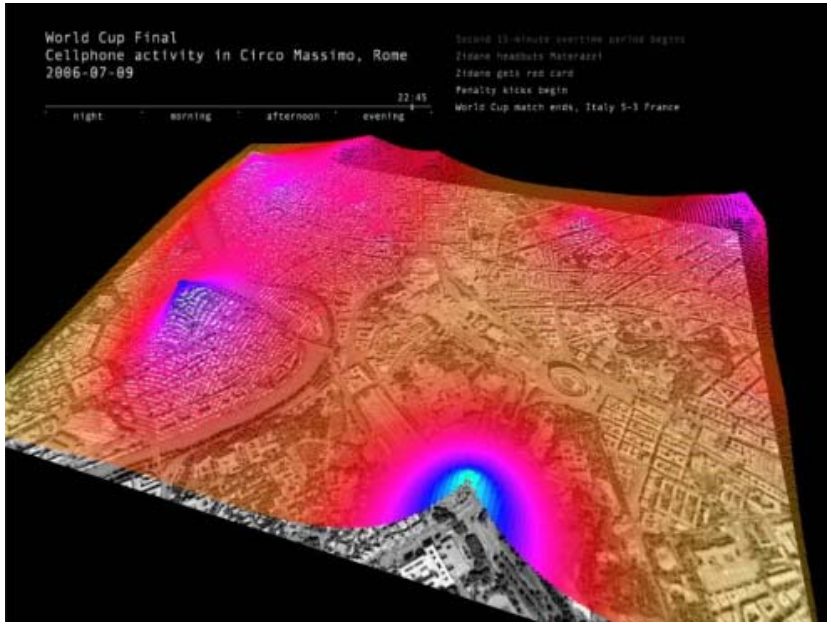
Cheng, Zhiyuan, et al. "Exploring Millions of Footprints in Location Sharing Services." *ICWSM* 2011 (2011): 81-88.

Social and Computer Science in the 21st century

New kinds of data (meso-scale)

Urban movement analysis from GPS/phone data

The Amsterdam Real Time Project

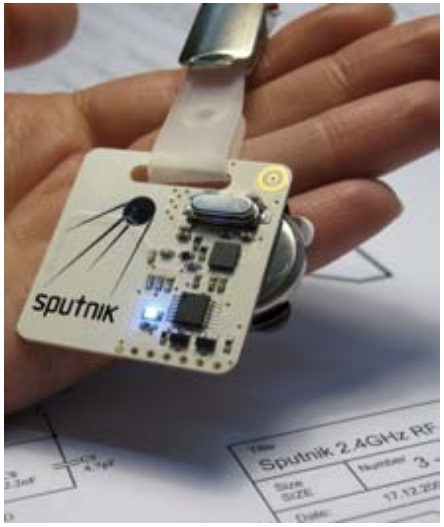


Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-time urban monitoring using cell phones: A case study in Rome, *IEEE Transactions on Intelligent Transportation Systems, 12*(1), 141-151.

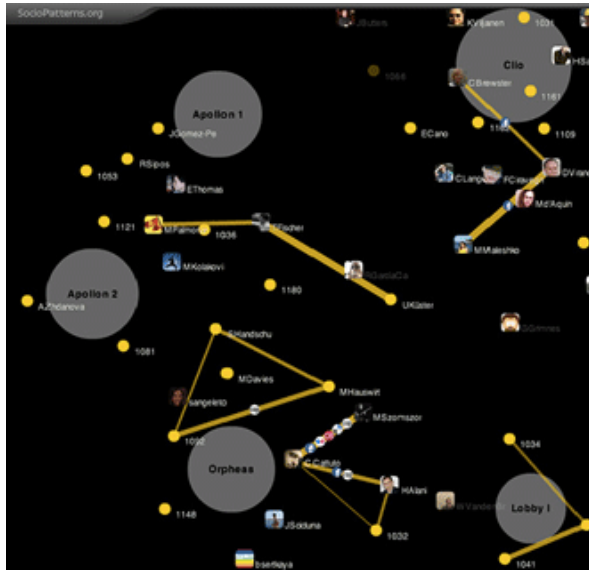
Social and Computer Science in the 21st century

New kinds of data (micro-scale)

Social Sensing via RFID

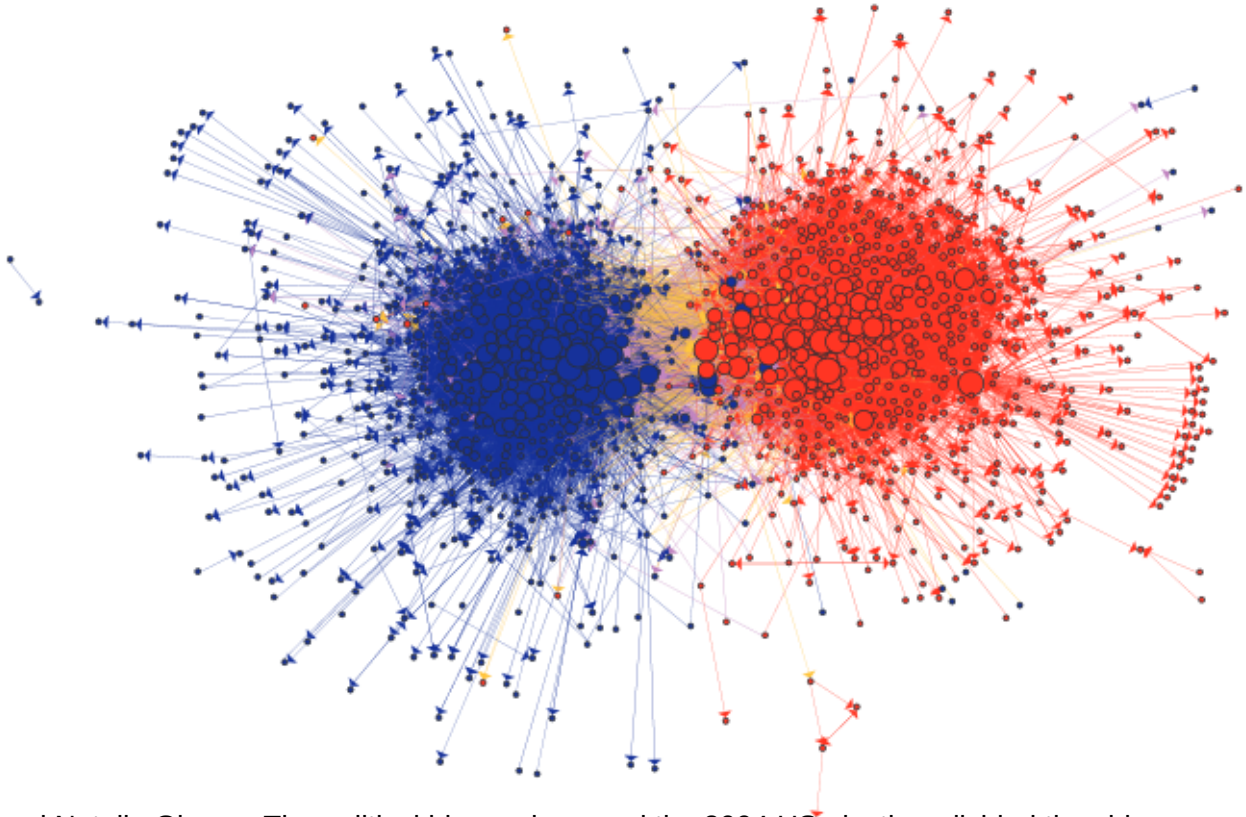


SocioPatterns



Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J. F., & Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PloS one*, *5*(7), e11596.

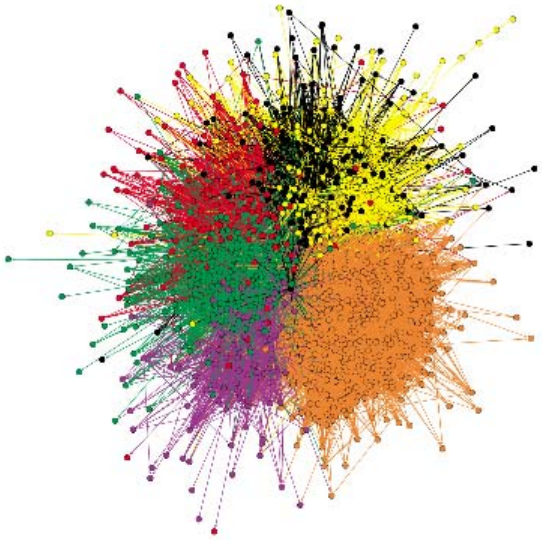
Assessing online conversational practices of political parties



Lada Adamic and Natalie Glance. The political blogosphere and the 2004 US election: divided they blog. *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005.

Assessing online conversational practices of political parties on Twitter

During the German National Election 2013

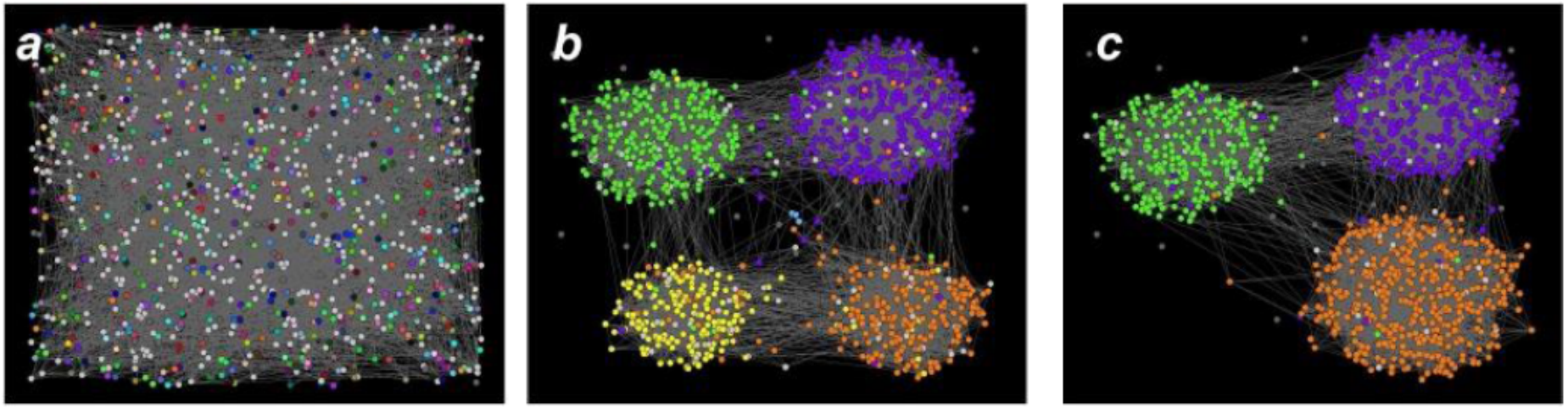


(a) Following (H=0.83)

Haiko Lietz, Claudia Wagner, Arnim Bleier, and Markus Strohmaier, When Politicians Talk: Assessing Online Conversational Practices of Political Parties on Twitter, The International AAAI Conference on Weblogs and Social Media (ICWSM2014), Ann Arbor, MI, US, 2014.

Computational Social Science

Evolution & the naming game



[Q. Lu, G. Korniss, B.K. Szymanski, The Naming Game on Social Networks: Community Formation and Consensus Engineering, Journal of Economic Interaction and Coordination, vol. 4(2), 2009.]

Inequality

Kay, Matthew, Cynthia Matuszek, and Sean A. Munson. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations."

Observations

The **digital world is tracking the social world more and more closely**.

This enables us to **use computation** to discover patterns, build models, validate social theories and learn about societies.

Which in turn **requires computational infrastructures** for dealing with:

More data: analyzing large amounts of data

Fuzzy data: cleaning up inprecise and noisy data

New kinds of data: processing real-time sensor streams and web data

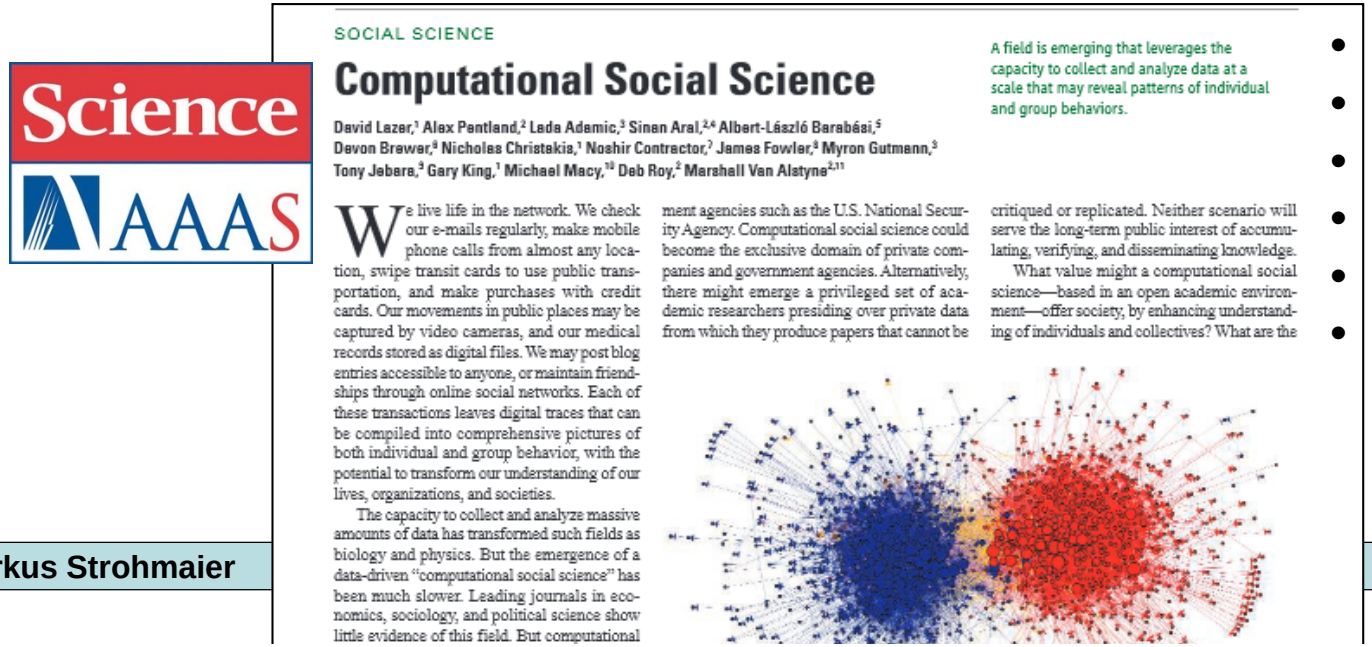
Correlations: understanding What (in addition to Why)

A lot of this data is on the Web!

Computational Social Science

Computational Social Science:
"The science that investigates social phenomena through the medium of computing and algorithmic data processing."
[adapted from CSSSA]

CSSSA: http://computationalsocialscience.org/



Science AAAS

SOCIAL SCIENCE

Computational Social Science

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[9] Tony Jebara,[2] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational ment agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

- Harvard iQS,
- Stanford IRiSS,
- CMU CASOS,
- ESRC COSMOS
- Web Observatories
- …

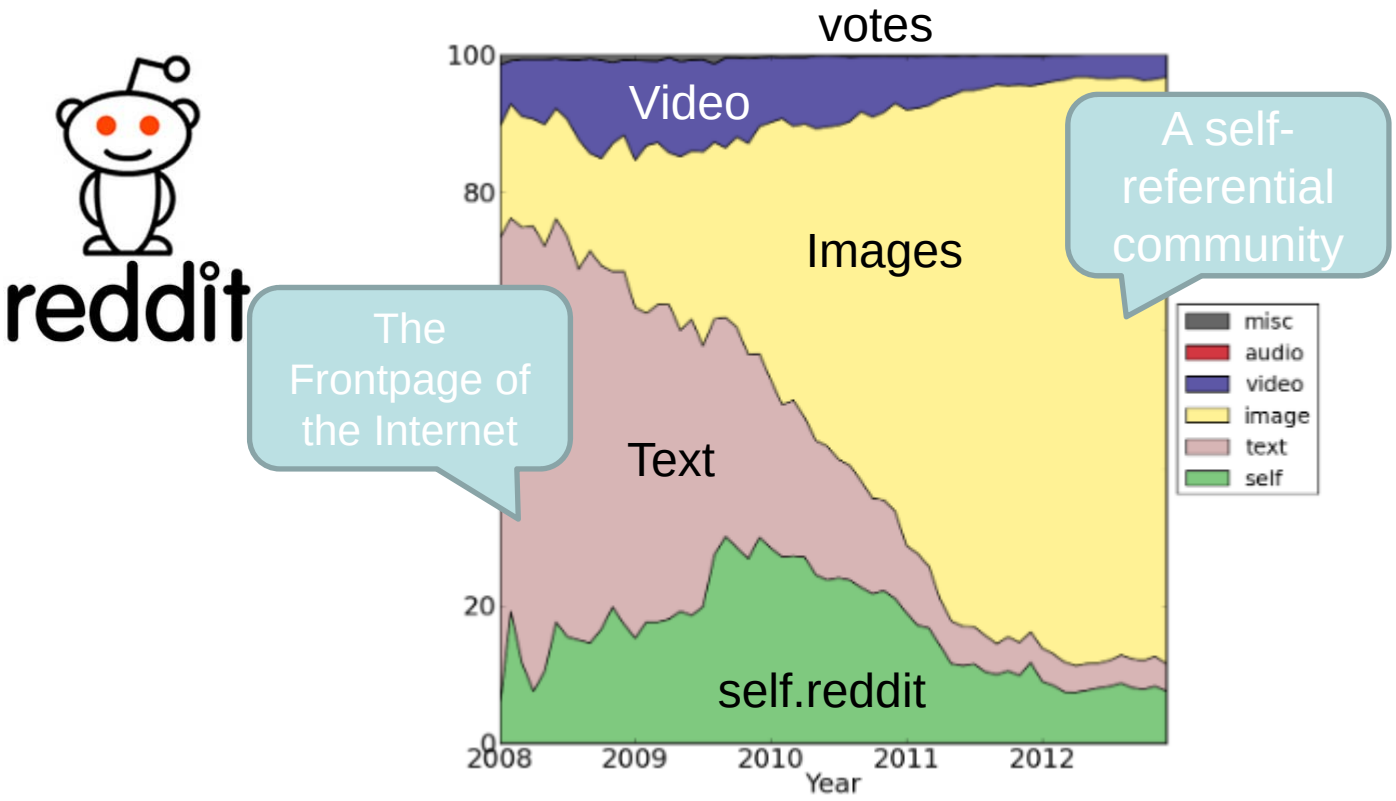
Random Social Connections

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/

Evolution of Online Communities



P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel and M. Strohmaier. Evolution of Reddit: From the Front Page of the Internet to a self-referential community? In 23rd International World Wide Web Conference (WWW2014), Web-Science Track, Seoul, Korea, 2014.

Observing political movements and conversations

Analysis of political conversations on Twitter (Egyptian revolution 2011)

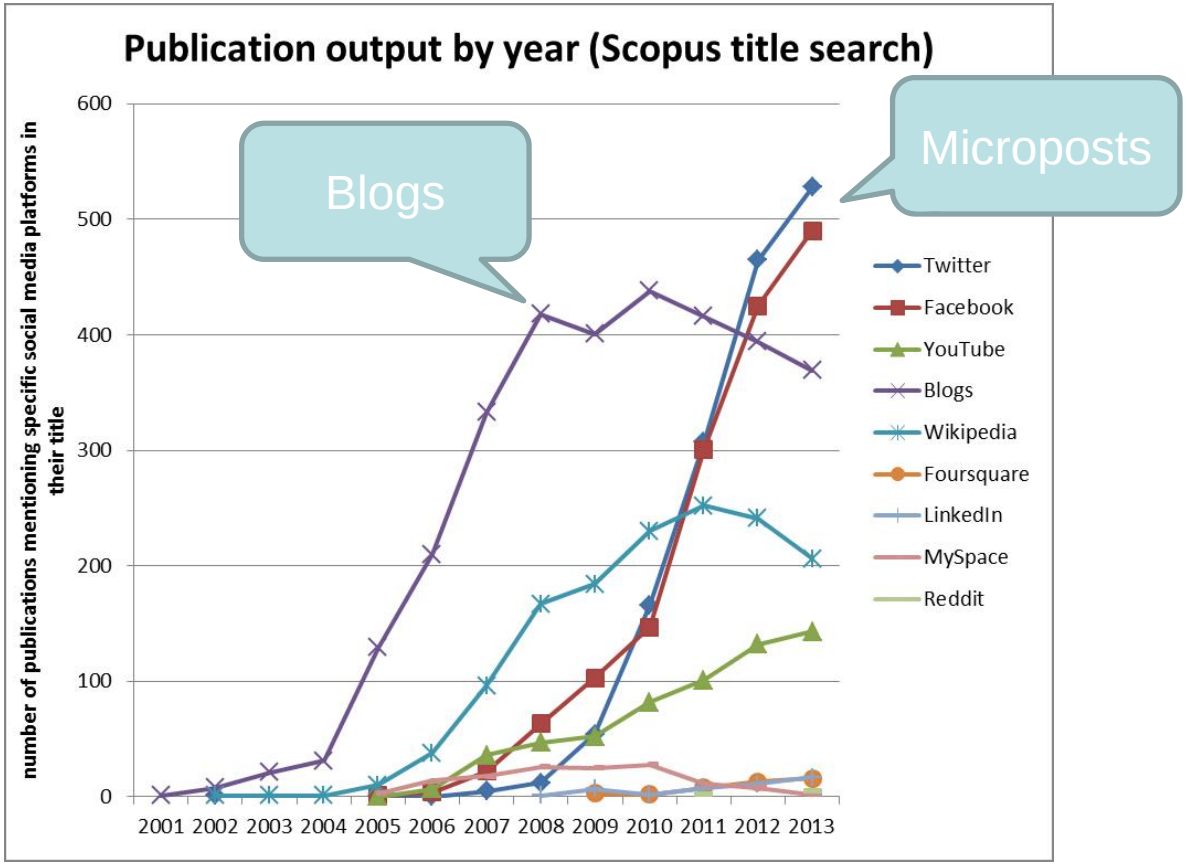
23.Jän	24.Jän	25.Jän
qaeda-linked	**#jan25**	**#jan25**
blames	blocked	blocked
militant	protests	**tahrir**
palestinian	bust	protest
al-qaida-linked	revolution	protest
group	anti-government	tear
army	berlin	thousands
bombing	prepares	#25jan
korean	nefertiti	demonstrations
#jan25	braces	violent
interior	#25jan	**#tahrir**
qaeda	demonstrations	anti-government
ceo	#arabprotest	square
al-qaeda-linked	activists	armored
church	ceo	riot
#palestinepaper	return	tiananmen
accusations	#j	demonstrators
attacked	kate	protestors
conclusive	tunisian-inspired	fired

1. day of the protests

Based on an analysis of about 100 mio tweets on egypt

Days on which the internet was shut down

And there is an increasing amount of computational social science research being conducted…



Weller, K. (2014). Bibliometric analysis of social media research: Publication output for different social media platforms. Blog post, 07.04.2014. Retrieved from: http://kwelle.wordpress.com/2014/04/07/bibliometric-analysis-of-social-media-research/

More examples

Predicting…

personality from twitter
Golbeck et al. 2011

depression via social media
De Choudhury et al. 2013

elections with Twitter
Tumasjan et al. 2010

crime using Twitter
Gerber 2014

stock market indicators
Zhang et al. 2011

flu trends using twitter data
Achrekar et al. 2011

box-office revenues
Asur & Huberman 2010

N=all

Online = Mirror of Offline

the social atom

Course team

- Instructors:
  - Markus Strohmaier
  - Claudia Wagner
  - Philipp Singer
  - Fabian Flöck




- For general, course- or assignment-related questions, please use the newsgroup

- NG name: infko.compsocsci
- NG URL: news://news.uni-koblenz.de/infko.compsocsci

About me

Background:

- April 2013 – present: Prof., U. of Koblenz & Scientific Director GESIS
- July 2007 – April 2013: Ass. Prof. (Univ.Ass.), TU Graz, Austria
- 2006 - 2007 15 months Post-Doc, University of Toronto, Canada
- 2002 - 2006 Researcher, Know-Center, Austria

Education:

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

About me

Research Background:

- Web Science / Knowledge Management

Research Interests:

- Web Science with a focus on networks and Social Computing
- Computational Social Science

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

- Course #04199, Computational Social Science
  - 2nd year at Koblenz
  - Still under development!
  - Home assignments
- Part of
  - WI, eGov, IM und Web-Science
  - **programming skills** are required
- Your feedback is appreciated

What are you studying?

Course Organization and Logistics

- ***Lectures***
  Wednesdays 16:00 – 17:30, (see KLIPS)
  Apr 2015 – July 2015,
  Room **E 524**

- ***Websites:***
  Klips: https://klips.uni-koblenz-landau.de/v/71114
  Web: https://gesiscss.github.io/csskoblenz/

- **Enroll!**
  In order to obtain a grade, you need to enroll for this course:
  3. Belegphase 17.04.2015 - 01.05.2015
  Abmeldephase 16.03.2015 - 17.05.2015

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 29.7.2015, no aids are allowed

- **Prerequisites**: **General Programming Knowledge**
-

In order to successfully complete the course, you need to have a combined 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

Submission of Home Assignments

2 Home assignments

- Individual work
- Based on Python
- Network analysis, modeling, generation

More information in an upcoming tutorial (next week, given by Philipp Singer).

Policies

- **Course documents**: Assignment descriptions and lecture notes will be made available via 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 checked for plagiarism.

If you submit plagiarized code on any home assignment, you will not be able to complete this course this year, and have to repeat next year.

(Incomplete) FAQ

Q: My colleague plagiarized my code, but I did all the work myself. Do both of us have to repeat the course again?

A: Yes, we can not verify who copied from whom. Sharing code is prohibited. Both students have to take the class again.

Q: I plagiarized just 1 of n home assignments. Does this mean I can not complete the entire course?

A: Correct. If you submit plagiarized code for any of the home assignments, it means you have to repeat the class next year.

Q: Does an automated program decide whether my code represents a case of plagiarism?

A: No. Each case will be reviewed by teaching assistants and lecturers before we make a decision.

Course Organization

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

- **discuss methods, theories and papers**
- **go through the theory necessary for home assignments**
- **answer questions related to home assignments**

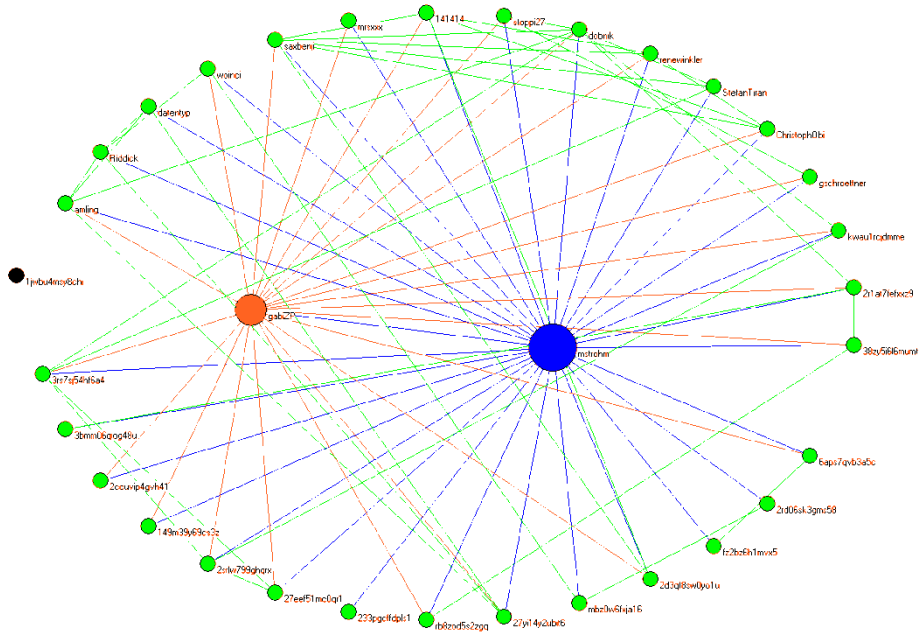
Course Goals

- To equip you with the basic **knowledge** and **tools** for performing independent computational social science based on data from the web
  - Knowledge:
    Social network theory, social network dynamics, social network evolution, communities, ….
  - Tools:
    Python, iPython, iPython notebooks, simulators, network metrics

Excerpt!

Home Assignments HA1-2

- Tools
  - Python
- Networks
  - One-Mode Networks
  - Two-Mode Networks
- Analysis
  - Lattices
  - Metrics, such as Centrality, Redundancy
  - Ranking
  - Network simulators
  - Network infection



How many of you know Python?

Home Assignment 1 will be handed out next week!

Home Assignments

- 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

- Found data
- What is social network theory? Why is it relevant for the web?
- How do networks evolve?
- How do you search in social networks?
- What are social parameters of networks?

How many of you know…

- 6 degrees of separation (small world problem)
- Power law networks
- Network generators
- Degree Distributions
- Galois Lattices
- …

Preliminary Course Schedule

WEEK	WHEN		WHAT	WHO
1	15.4.		CSS Introduction & non-obtrusive measures	Markus
2	22.4.		Sociology: Small World Networks & Properties	Markus
		Ü	Python Tutorial, HA 1 Vorstellung	Philipp
3	29.4.		Sociology: Strength of Weak Ties	Markus
4	6.5.		Sociology: Prestige and Communities in Social Networks	Markus
5	13.5.		Sociology: Affiliation Networks and Galois Lattices	Markus
6	20.5.		Python Stats Tutorial, HA 2 Vorstellung	Claudia
7	**27.5.**		**PFINGSTEN**	
8	3.6.		Übungen	
		Ü	Abgabegespräche, Python Scikit Learn Tutorial (3 Stunden)	Philipp, Claudia
9	10.6.		Anthropology / Cultural studies	Fabian
10	17.6.		Pol. Science: Online political discourse	Markus
11	24.6.		Soc. Psych. and Soc. Econ: Language Games	Markus
12	1.7.		Soc. Psych. and Soc. Econ: Networks Dynamics 1	Claudia
13	8.7.		Soc. Psych. and Soc. Econ: Networks Dynamics 2	Claudia
14	15.7.		Research Methods CSS 1	Claudia
15	22.7.		Research Methods CSS 2 (eventuell - sonst invited talk)	Claudia
		Ü	Abgabegespräche	
	29.7.Prüfung			

presence in class

Non-Goals

In the research community, there is **no consensus** regarding the theoretical foundations of „Computational Social Science“ yet

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

Instead of giving an authorative account of CSS, this course aims to give an overview of **prominent**, **interesting** and/or **powerful research results** generated by related fields so far.

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!

- Computational Social Science-

Based on the following resources

Unobtrusive Measures, Eugene J. Webb, Lee Sechrest und Donald T. Campbell

Raymond Lee (2000). Unobtrusive Methods In Social Research, Buckingham: Open University Press.

Allan Kellehear (1993). The Unobtrusive Researcher - A guide to methods, St Leonards: Allen & Unwin.

William G. Zikmund, Barry J. Babin, Jon C. Carr (^{9}2012). Business research methods. Cengage Learning (chapter 11)

Yoram M Kalman. Unobtrusive Methods for Social Science Research - A Neglected Methodological Approach in the Social Sciences (slides)

Christine Hine (2011). Internet Research and Unobtrusive Methods. *Social Research Update*, *61*, 1-4.

Michelle O'Brien (2010). Unobtrusive Research Methods – An Interpretative Essay.

Methods of data capture in social research

Surveys
Interviews
Experiments
Observation
Document analysis
Artefact analysis
…

Types of physical traces

There are two general types of trace measures:

Accretion - a build-up of physical traces

Erosion - the wearing away of material

Physical Traces: Accretion

Before and after electric waste disposal units: change in fly population measured on car grilles

Graffiti:

- Racial tensions in Hawaii
- Youth relationships in high schools
- Correctional facility for male juveniles
- The writing on desks by classroom/grade level

Garbology:

- Survey data on alcohol consumption vs. number of empty alcohol containers appearing in garbage cans. The survey data underestimated consumption due to high refusal rate and exclusion of teenage drinkers
- Others: Condom wrappers 1976-1984, and beyond; Beer tabs vs. beer cans as evidence for recycling.
- In US, 4th amendment issues

Nose prints on exhibition cabinets

Physical data – Erosion

- Floor tiles around chick exhibit
- Popularity of gym apparati by chalk consumed
- Popularity of library books by smudges, finger marks etc.
- Leaflets in different languages in tourist sites
- Postcards in museum exhibits
- Paper tissue and cough medicine in campus store correlated with class attendance records

Found Data – Disadvantages

- Conservative estimate
- Socially dependent
- Takes time to accumulate
- Inferentially weak
- Privacy issues

Retrieved Data – Running Records

Running records (vs. episodic and private records) are:
e.g. mass media, reference works, records of proceedings

Advantages
- cover lengthy periods of time
- ubiquitous
- lower cost
- less amenable to self-report bias
- allow the exploration of trends and temporal patterns

Limitation: collected by others

Retrieved Data – Personal and Episodic Records

Best example is personal documents:

- Letters
- Diaries/ daily journals
- Autobiographies
- CV’s
- Wills
- Photo albums

Retrieved Data – Disadvantages

Often can't be used "as is". Should consult with those who produced them (reactivity!)

Quality issues, especially when more extensive

Selectivity: exclusion and inclusion criteria

Statistics might reflect more the organization collecting the data, than the sources of the data

Confidentiality

Captured Data

Exterior physical signs: head and facial hair, tattooing, clothing and adornments

Expressive movement: demeanor, eye gaze, touching, verbal latency

Physical location: proxemics, spatial arrangement

In-situ conversation: pronunciation

Time related behavior: often neglected. Duration as proxy for importance, time of day influencing behavior

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

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

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