monitoring socio-technical adaptive networks

david m.s. rodrigues
eth zürich, february 28, 2011

Lisbon, Kings Palace Square - Dirk Stoop, 1662

first appeared in Camille Flammarion's L'atmosphère: météorologie populaire (Paris, 1888)

Paul Butler, Facebook's social graph, Dec. 2010

what is relevant in the age of zettabyte (2^{70})?

- According to the annual survey of the global digital output by International Data Corporation, the total amount of global data was expected to pass **1.2 zettabytes sometime during 2010**.

- Mark Liberman calculated the storage requirements for all **human speech** ever spoken at **42 zettabytes if digitized as 16 kHz 16-bit audio**.

- Research from the University of Southern California reports that in 2007, humankind successfully **sent 1.9 zettabytes** of information through broadcast technology such as **televisions** and **GPS**.

- Research from the University of California, San Diego reports that in 2008, **Americans consumed 3.6 zettabytes** of information.

http://en.wikipedia.org/wiki/Zettabyte

what is relevant in the age of zettabyte (2^{70})?

- **Challenge**: understand how and what are people discussing, what are their interests, subjects, opinions, arguments, communication structure

- **Goal**: characterise opinion dynamics extracted and deduced from large and diverse data

four basic ideas
real-time, relations, knowledge and dynamic

Real-time monitoring of multi-level **network structures** for the study of **knowledge** generation and **opinion dynamics** in the Internet

four foundation areas for action

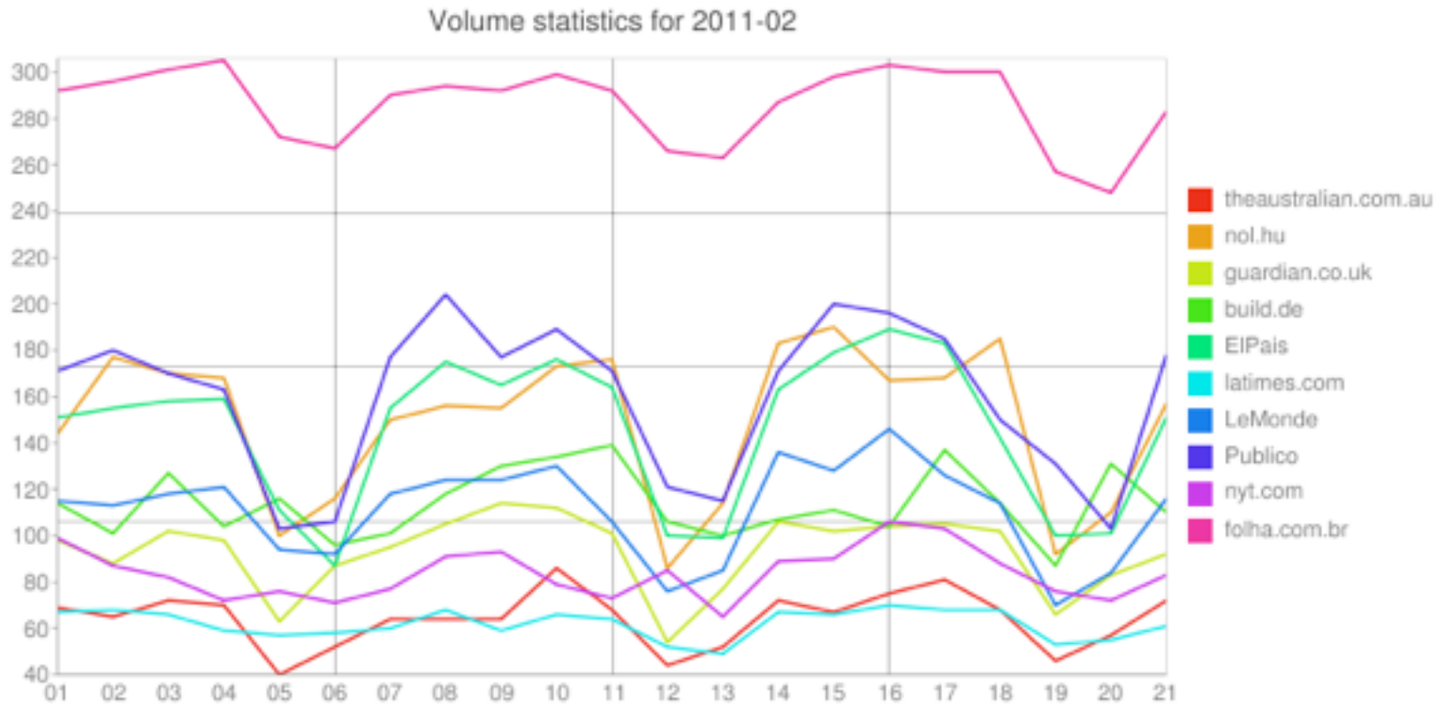
realizations in the observatorium

empirical data
collecting newspapers for 15+ months

- (1yr+) News (10 world general newspapers) 27,9GB (of text)-> 424.850 news (as in Feb 17th)

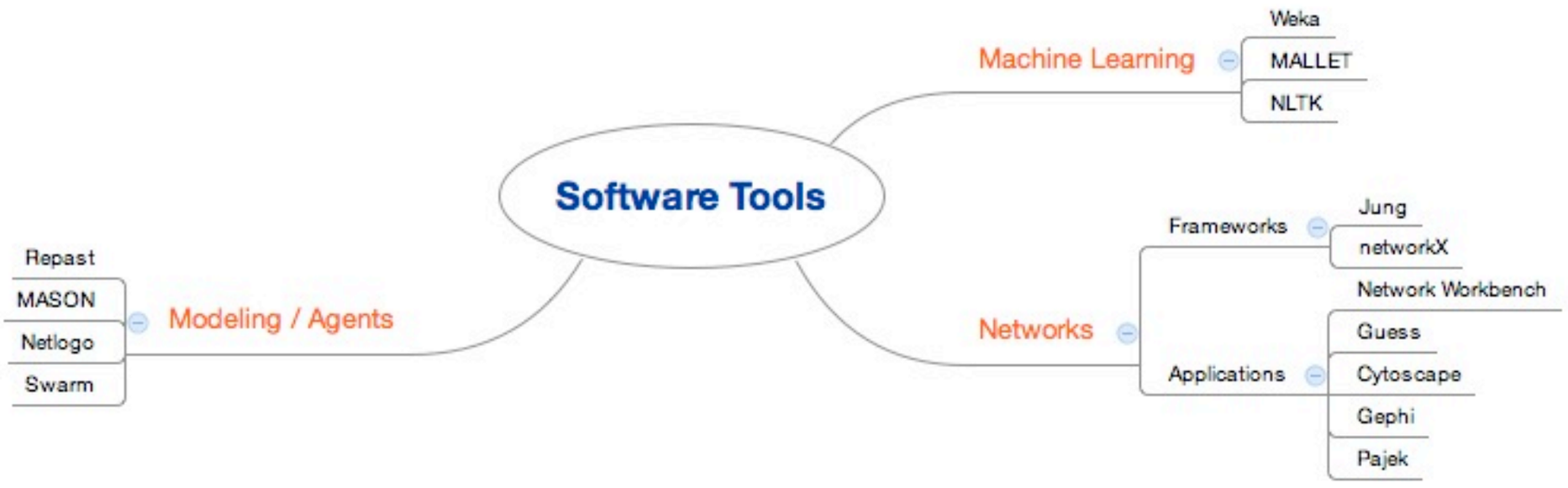
- (3m+) Recently started collecting financial news from 50 online financial newspapers

- (5m+) Collecting Twitter data



tools
why develop?

- There are many tools available.
  - Exercise in development helps to understand the inner mechanisms of the theories
  - Allows tight control and integration with other software that we develop.
  - Small team doesn't use everything and therefore it's not necessary to invest time and scarce resources on learning software packages



tools & infrastructure
theseus: a library for topic extraction in python

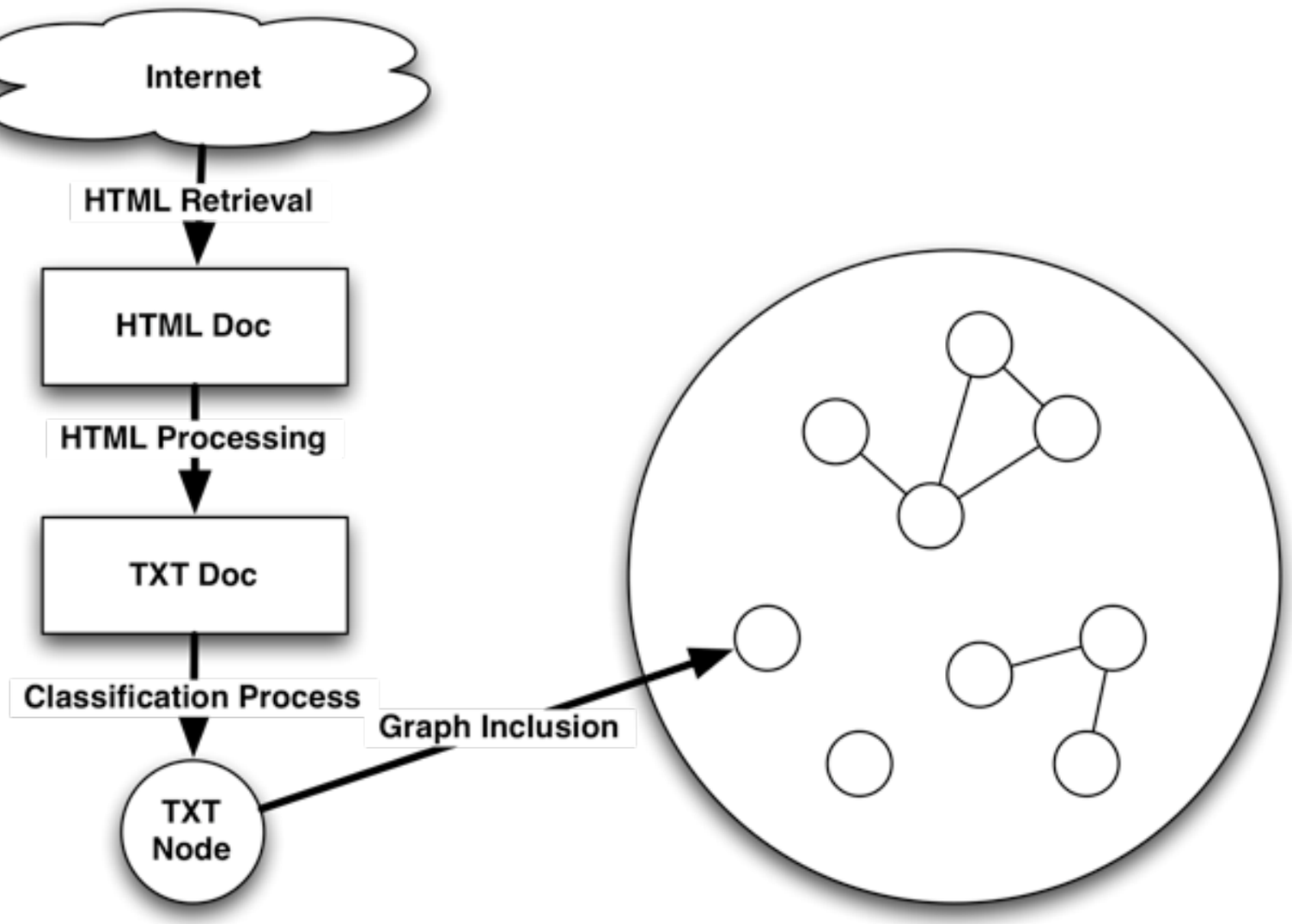
- **theseus** - a python library to automate topic extraction, summarization and visualization with guess.
- not yet production ready (0.7.1)
- want to contribute?

  git://github.com/sixhat/theseus.git

http://www.theobservatorium.eu/html/

algorithms
information theory in topic extraction algorithm

- Variation of Information on a sliding window of networks of news items to extract cluster of topics.



algorithms
information theory in topic extraction algorithm

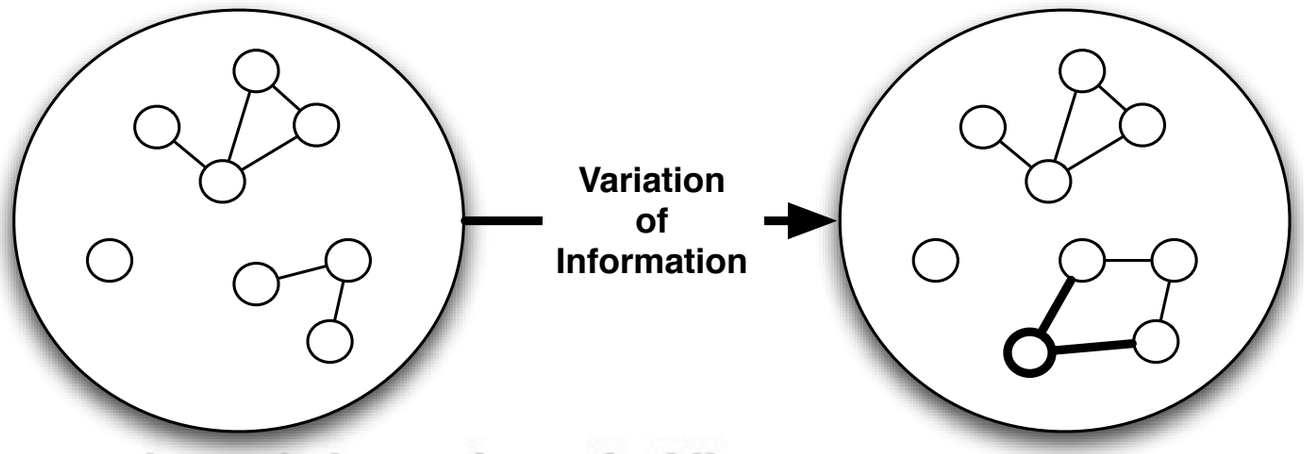
- Variation of Information on a sliding window of networks of news items to extract cluster of topics.

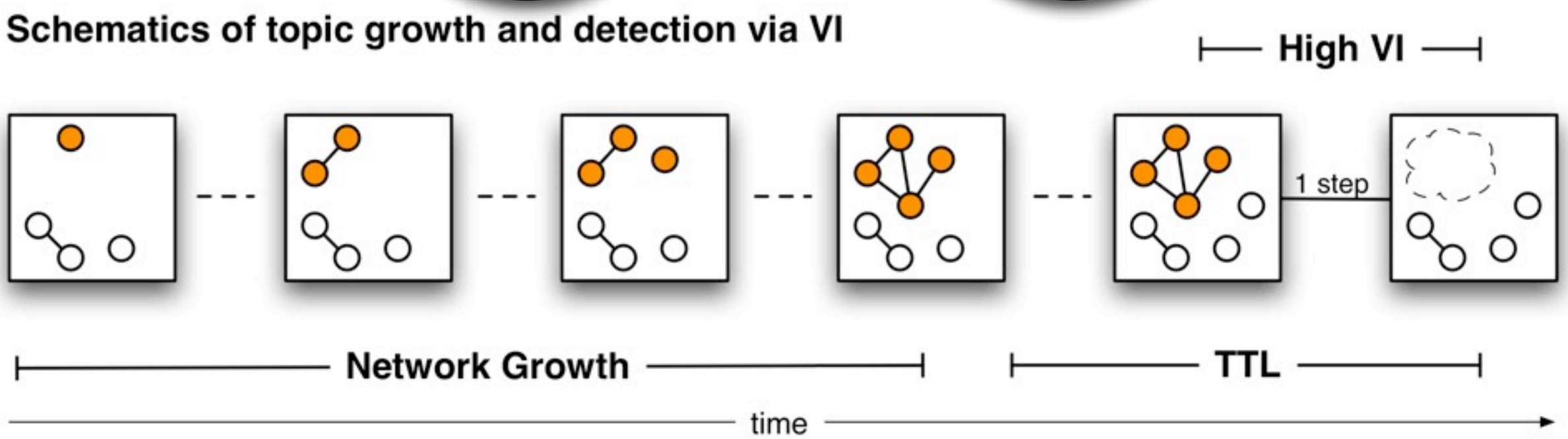
  - **Edge Creation** (Similarity between nodes)

    - Jaccard similarity $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$
    - Cosine similarity
    - tf.idf *(with the corpus being either the documents in the window or a pre-computed corpus)*
    - (...)

algorithms
information theory in topic extraction algorithm

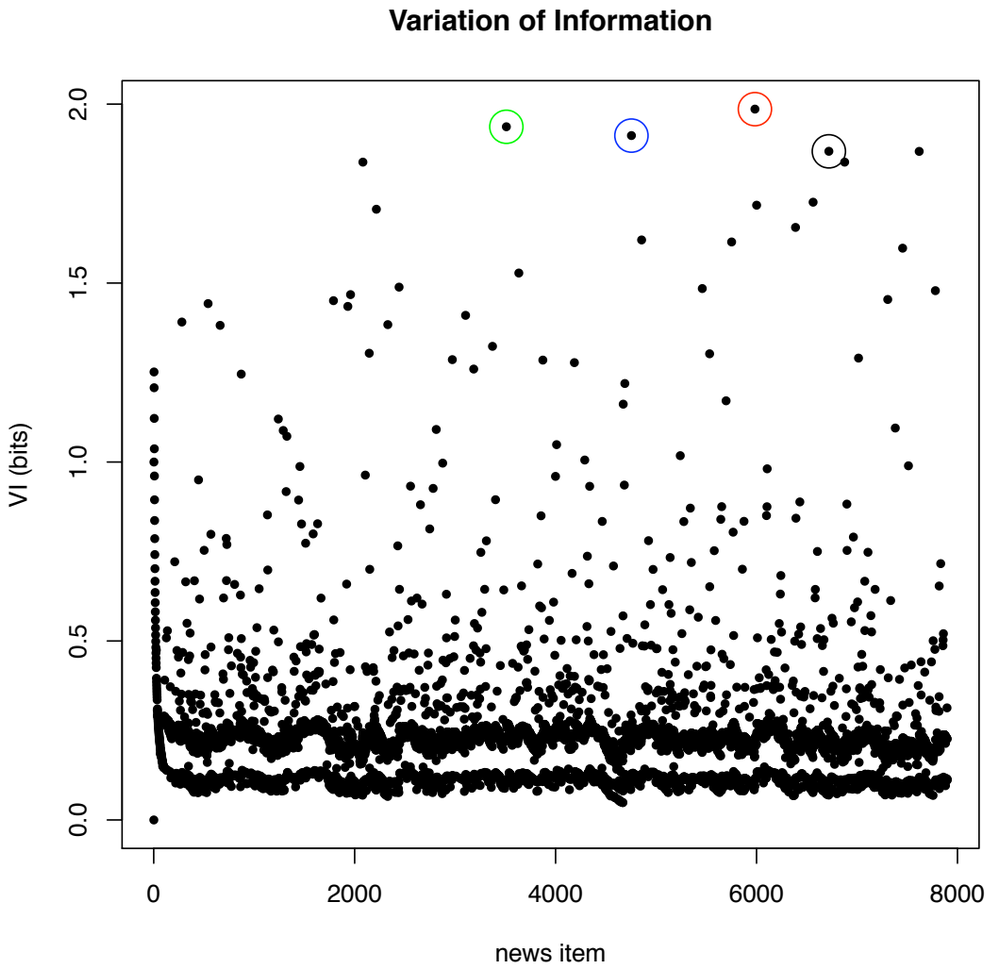
- Variation of Information on a sliding window of networks of news items to extract cluster of topics.

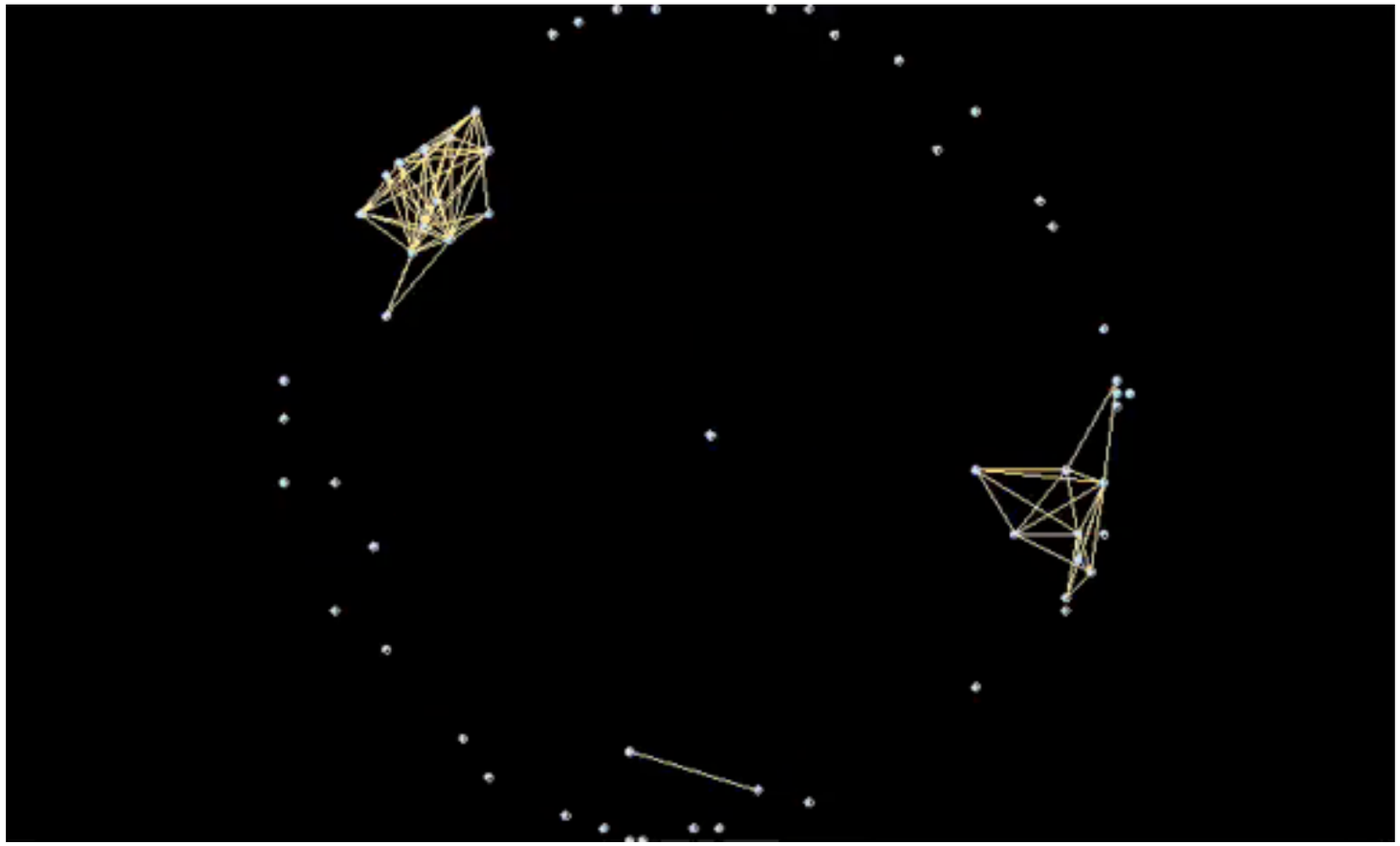




algorithms
information theory in topic extraction algorithm

- Variation of Information on a sliding window of networks of news items to extract cluster of topics.





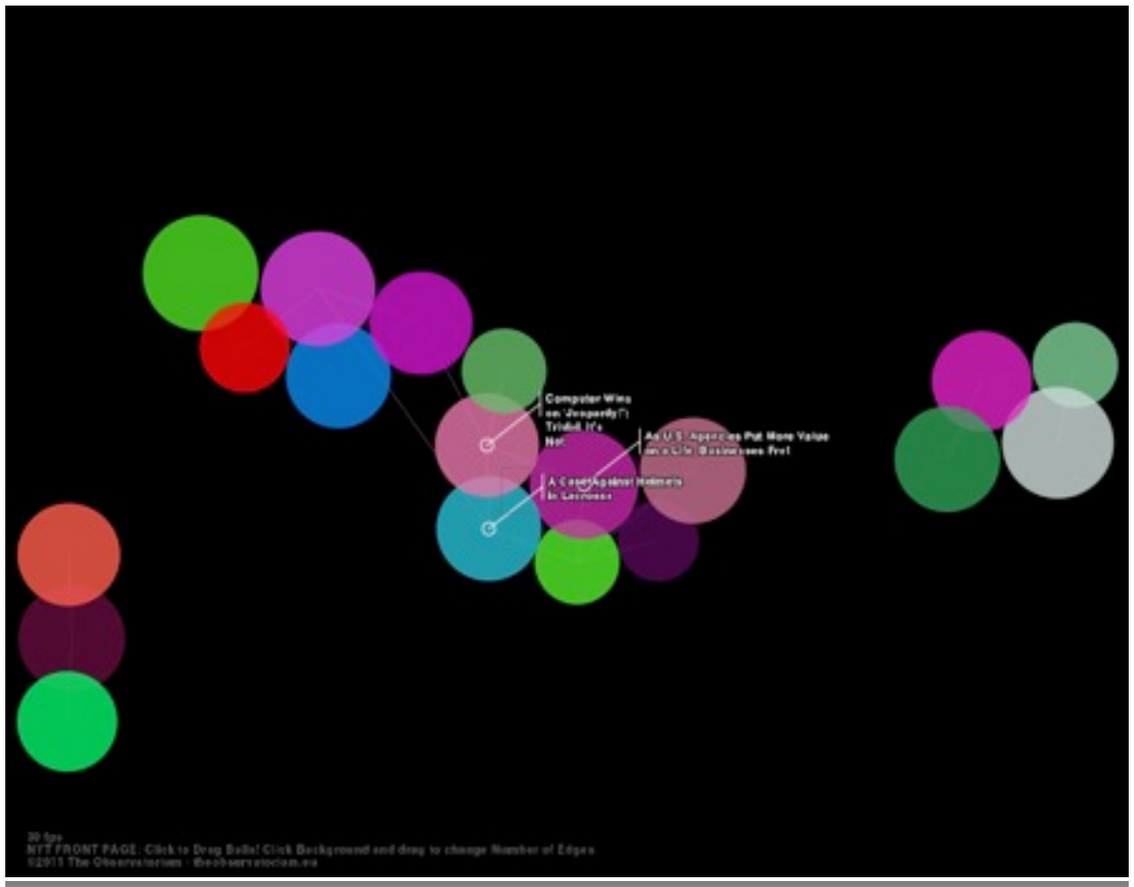
5x

visualization and human-computer interaction

- Online display of information in real time (or close to...)
    - Web - now using Java Applets, but want to move to full HTML5.
    - Other - using ‘computer-less’ presentations of data with everyday technology
        - Microsoft Kinect - create ‘controller-less’ interactions with data

demo

exploring the new york times front page



http://theobservatorium.eu/nyt-demo/

exploring computer science @ ArXiv

http://theobservatorium.eu/arxiv-cs/

http://theobservatorium.eu/arxiv-cs2/

my research interests

2009 - informal communication networks

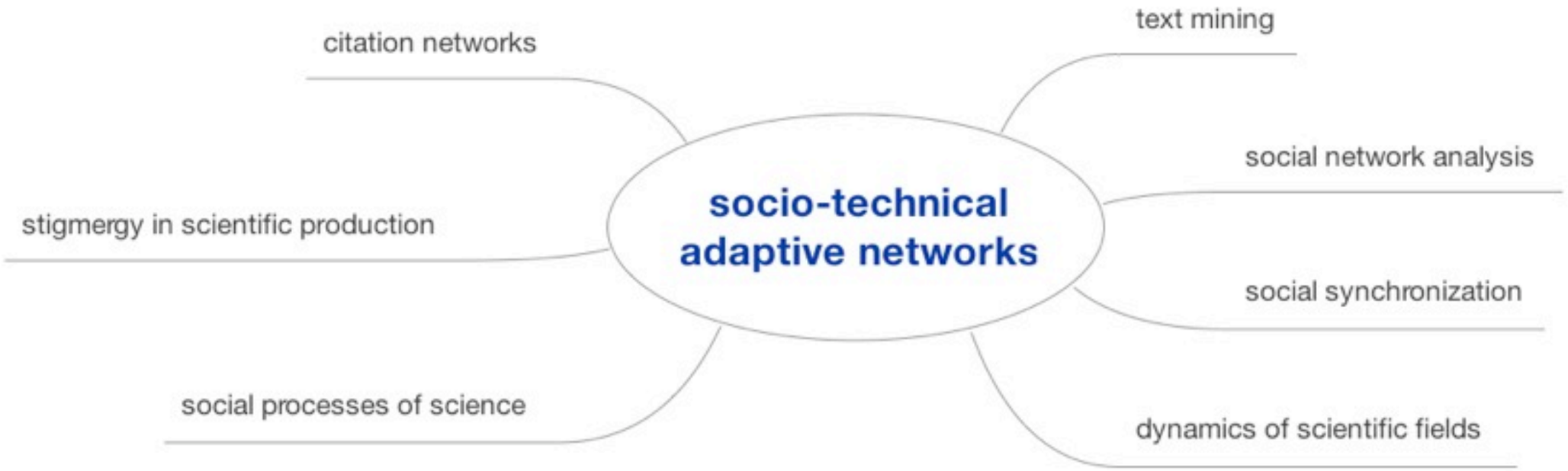
- 2007 - got interested in Complex Systems - started a MSc.
- Interested in emergence, agents etc...
- Mixing agent based simulation with real empirical data for community detection of informal communication networks (University email)
- Using Information Theory to track meaningful changes in topologies

- *pre 2007 - Chemical Engineer - Research on nanofiltration membranes for clavulanic acid broths*

2010/11 - information theory and synchronization

- Variation of Information to track topics in online media
- Synchronization phenomena on Adaptive Networks (mainly social)
    - Public Opinion Formation
    - Crowd-sourcing
    - Collective Intelligence
    - Stigmergy, collective behavior, herd behavior

social synchronization phenomena and the innovation accelerator



road ahead

- until the 2nd half of 2010 the observatorium was a team of... well... 2!
- we are now 6 and therefore more work can be produced
  - Politics http://work.theobservatorium.eu/presid2011/
  - Science http://work.theobservatorium.eu/science/
  - News http://work.theobservatorium.eu/news/
  - Finance (Tracking more than 50 financial newspappers from 20 top stock market countries)
- bootstrap the Social Theories foundation of the observatorium.

the observatorium is a long term project

- An no ending recollection of facts about the world trying to establish relations between those facts, actors and history

- My hope is that in the end we can understand our society better and try to paint pictures of it with the level of understanding similar to that of Dirk Stoop in the XVII century.

some references of our work

- Symons, J., Louçã, J., Morais, A., & Rodrigues, D. (2007). Detecting emergence in the interplay of networks. AAAI 2007.
- Louçã, J., Symons, J., Rodrigues, D., & Morais, A. (2007). Emergence in Social Networks: Modeling the Intentional Properties of Multi-Agent Systems. Proceedings of the 4th Conference of the European Social Simulation Association (ESSA'07) (pp. 639-650). Toulouse, France: Proceedings of the 4th Conference of the European Social Simulation Association (ESSA'07).
- Louçã, J., Rodrigues, D., Morais, A., & Symons, J. (2007). Pattern-oriented analysis of communication flow: The case study of cicada barbara lusitanica. In I. Zelinka, Z. Oplatková, & A. Orsoni (Eds.), 21st EUROPEAN Conference on Modelling and Simulation ECMS 2007 (pp. 229-234). Prague.
- Urbano, P., & Rodrigues, D. (2008). Rule based systems applied to online surveys. Proceedings of the IADIS WWW/ Internet 2008 Conference. Frieburg.
- Rodrigues, D., & Louçã, J. (2009). Mutual information to assess structural properties in dynamic networks. Euroepan Conference on Complex Systems 2009 (ECCS'09) Proceedings (pp. 1-15). Warwick.
- Rodrigues, D. (2010). The Observatorium - The structure of news: topic monitoring in online media with mutual information. Proceedings of the European Conference on Complex Systems 2010 (ECCS'10).
- Rodrigues, D. M. S. (2011). Multi-Agent-Based Simulation of University Email Communities. In G. Trajkovski (Ed.), Developments in Intelligent Agent Technologies and Multi-Agent Systems: Concepts and Applications (pp. 273-291). IGI Global. Retrieved January 5, 2011, from http://www.igi-global.com/Bookstore/Chapter.aspx?TitleId=49368.
- Louçã, J., & Rodrigues, D. M. S. (2011). Observation et analyse de réseaux de communication a grande échelle (Vol. 1, pp. 1-12).