

Challenges in analyzing online social networks

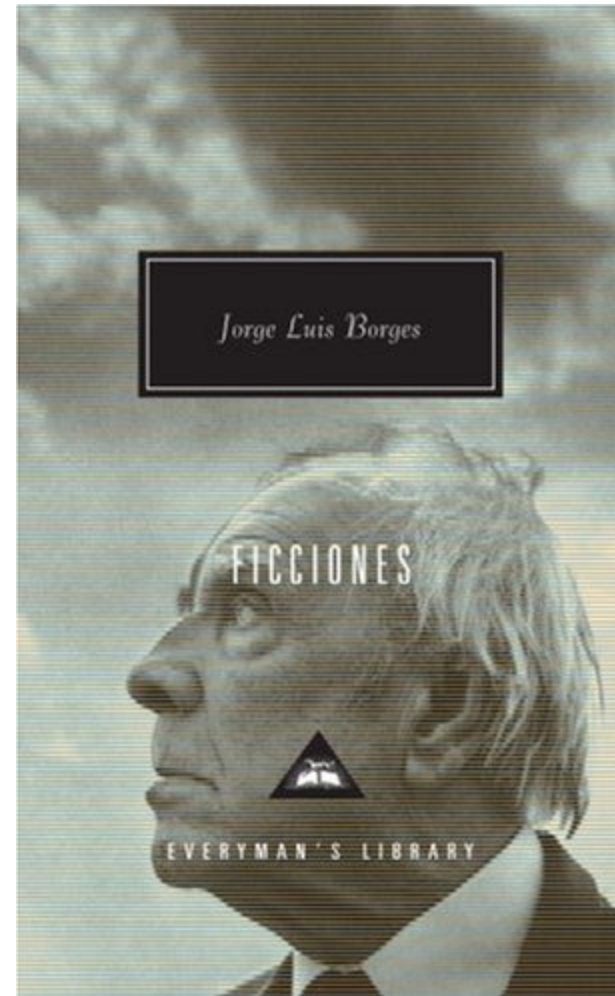
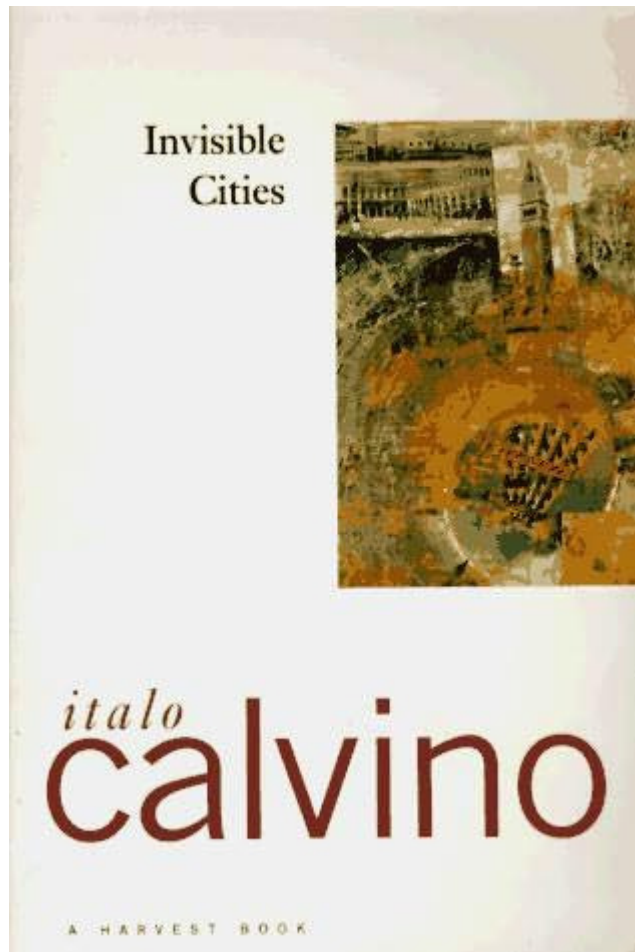
Virgilio F. Almeida
www.dcc.ufmg.br/~virgilio

Cambridge - December 2007



Computer Science Department
Federal University of Minas Gerais
Brazil

Online Social Networks: literature as a source of inspiration



Virgilio Almeida, UFMG 2007

Online Social Networks and the literature

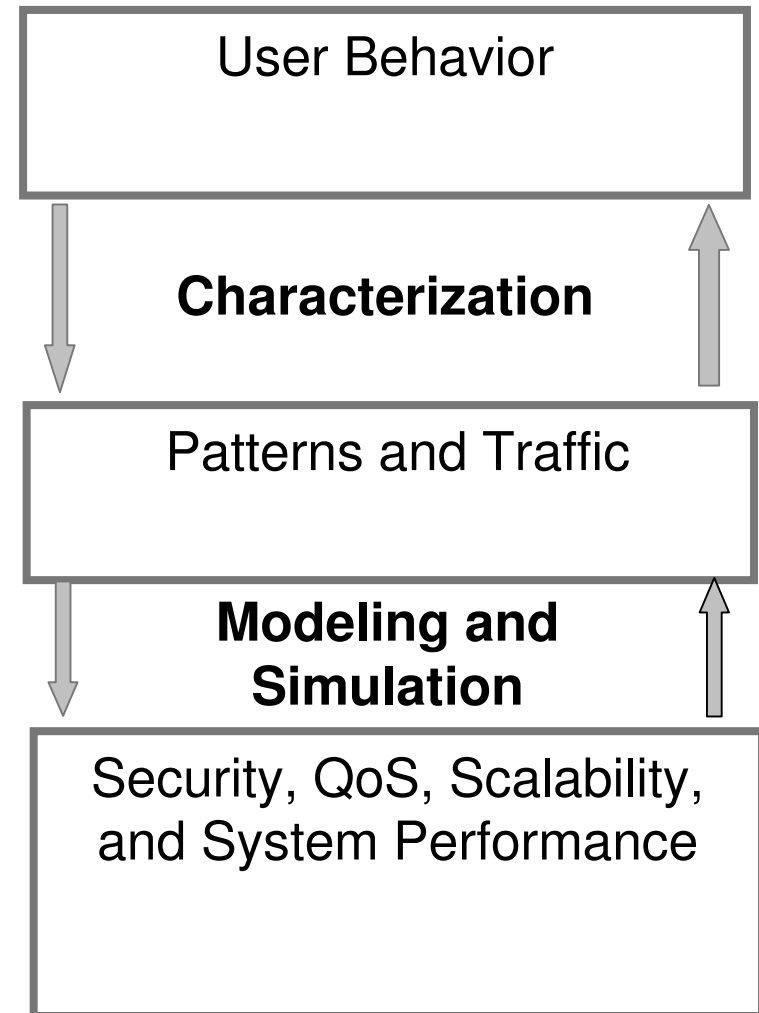
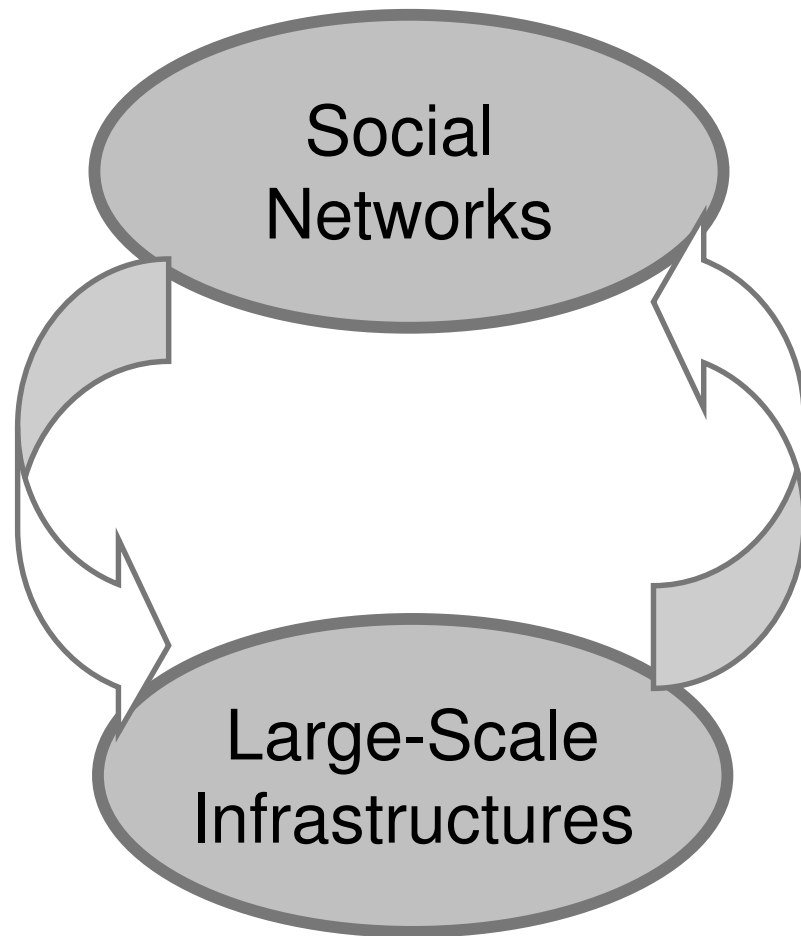
- "In Ersilia, to establish the relationships that sustain the city's life, the inhabitants stretch strings from the corners of the houses, white or black or gray or black-and-white according to whether they mark a relationship of blood, of trade, authority, agency“”.

Italo Calvino, *Invisible Cities*, 1972

- The impious maintain that nonsense is normal in the Library and that the reasonable (and even humble and pure coherence) is an almost miraculous exception. They speak (I know) of the “feverish Library whose chance volumes are constantly in danger of changing into others and affirm, negate and confuse everything like a delirious divinity.”

Jorge Luis Borges: *The Library of Babel*, 1941

Online Social Network: measuring and modeling



Challenges

- Need of measurement data
 - To observe and understand user behavior and content
 - Sampling large online social networks: representativeness
 - Privacy issues in data sharing
- Incomplete view of online social networks
- From populations to individual information
 - setting up algorithm thresholds
- Emergent properties (e.g., small world, preferential attachment)
- Tools and artifacts to test new algorithms and features for OSN:
 - Synthetic workloads
 - Simulation environment for large OSN

Some observations

- Social and technological networks are intertwined: email, IM, blogging, MySpace, Youtube,...
- Systems are able to collect social data at unparalleled scale and resolution;
- Online social networks measurements efforts have struggled with:
 - Size: there is too much to measure;
 - Single point of data collection: crawlers have to comply with robot rules and not engage in any activity that interferes with or disrupts OSN services;
 - Dynamics: workloads change fast.

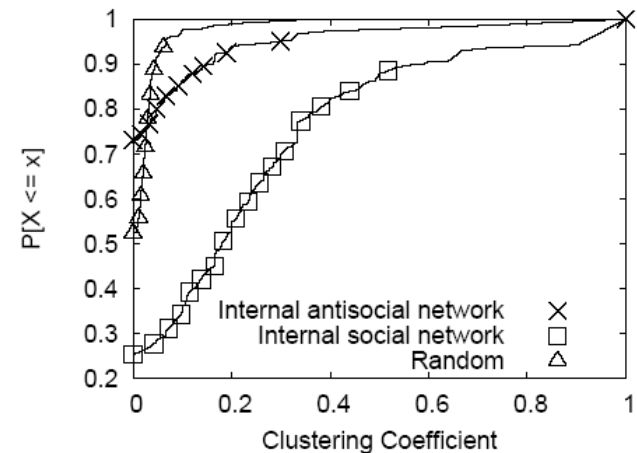
OSN data spans orders of magnitude

Small datasets → Massive datasets

- Citation network with 1736 nodes, actor collaboration with 392340 nodes... [Barabasi et al. 2005]
- Email network with 59812 nodes with emails of 5165 students [Ebel et al. 2002]
- Time evolution of a social network comprising 43,553 students. [Kossinets and Watts, 2006]
- YouTube 1.6 million-node, Flickr 1.8 million-node, LiveJournal 5.2 million node and Orkut 3 million-node [Cha 2007, Mislove 2007]
- 33 million blog requests to 210,738 blogs in a blogosphere [Almeida 2007]
- 30 billion of conversations among 240 million people: network of all IM communication over one month on Microsoft Instant Messenger [Leskovec and Horvitz 2007]

Measuring OSNs

- Sometimes incomplete but useful...
 - Power laws [Cha 2007, Mislove 2007]
 - Link symmetry [Mislove 2007]
 - Network properties of anti-social networks [Almeida 2005]
 - blogs[*]

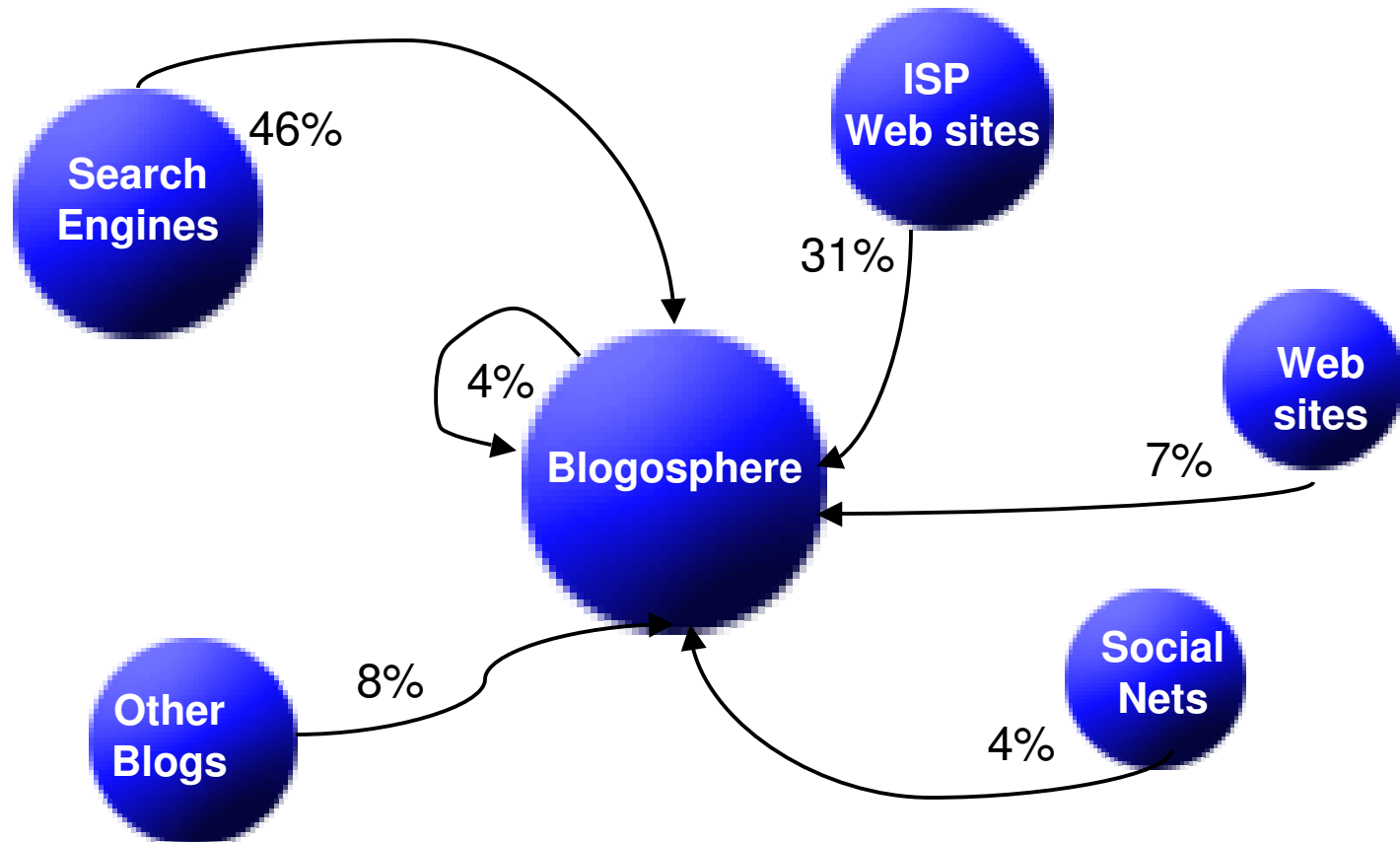


Web vs online social networks

- Web pagelinks arise out of the interactions of multiple individuals whose decisions are largely unconstrained by institutional rules and uncoordinated by procedures from central authorities/owners.
 - OSN have a single controlling entity with well defined interactions
- Design new systems and improve the existing ones
 - Malicious behavior in online social networks: content pollution, spam, self-promotion, etc.

Blogs: Origin of User Sessions

- Fraction of user sessions that used search engines and web links to enter into the blogosphere



- Traffic Characteristics and Communications Patterns in Blogosphere, ICWSM07, Almeida, Bestavros, Duarte and Almeida.*

OSN: characteristics

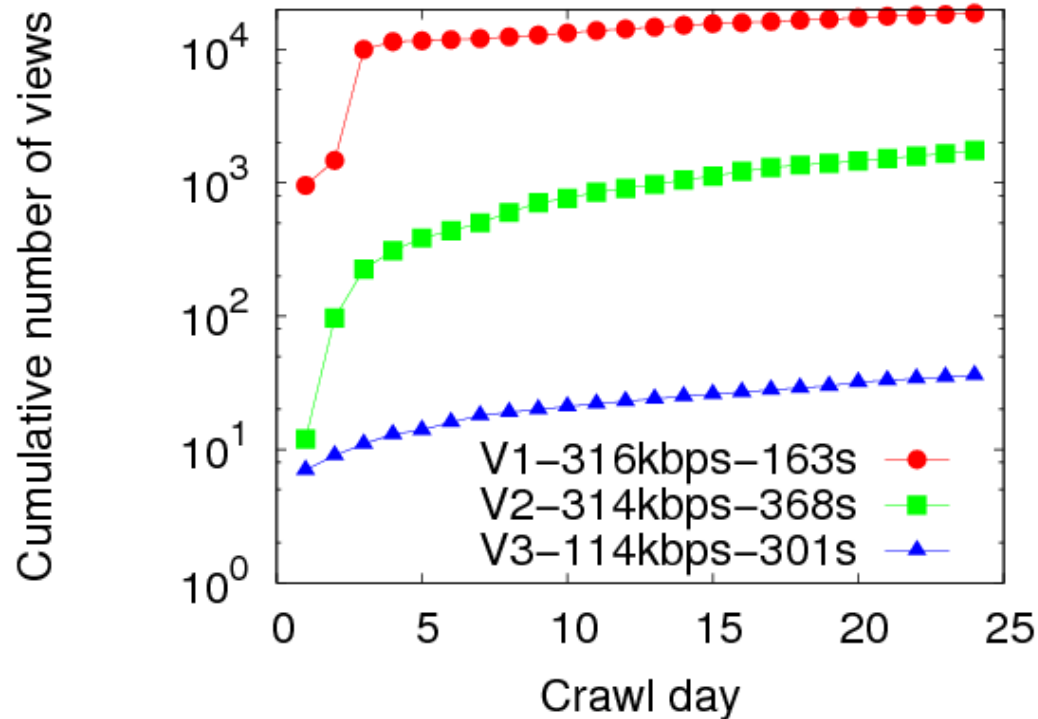
User generated content

- “This finding made it possible, three hundred years ago, to formulate a general theory of the Library and solve satisfactorily the problem which no conjecture had deciphered: **the formless and chaotic nature of almost all the books.**”
- “...every copy is unique, irreplaceable, but (since the Library is total) there are always several hundred thousand imperfect facsimiles: works which differ only in a letter or a comma. “

The Library of Babel, Borges, 1941

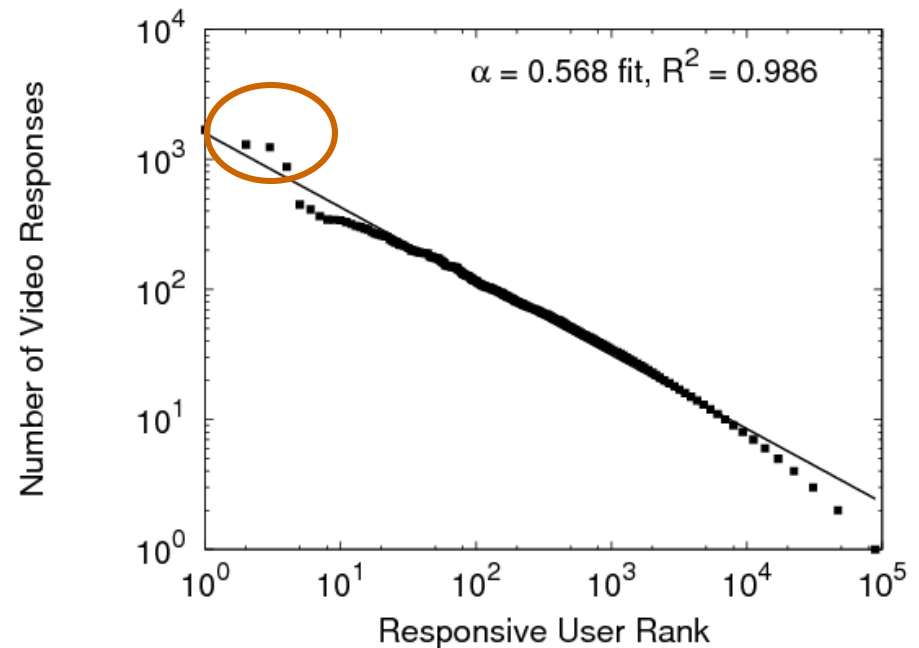
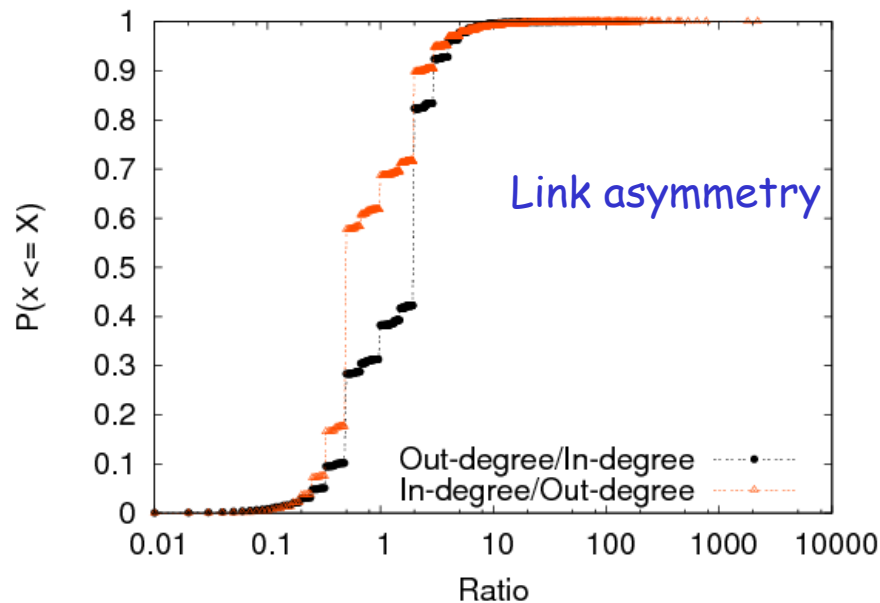
Online Social Networks

- Diversity and changes in user generated content



OSN: malicious behavior

- “There are also letters on the spine of each book; these letters do not indicate or prefigure what the pages will say. I know that this incoherence at one time seemed mysterious...”



OSN: conclusions

- We are now able to observe real-time interactions of millions of people at many different resolutions, based on access to electronic data sets such as e-mail records, mobile phone call logs, interaction of million of individuals in OSNs, but there are enormous challenges...
- Strategy to measure online social networks
 - It is not doable to measure everything we want
 - What should our measurement strategy for current OSNs be? How to measure a decentralized OSN?
 - Many current empirical studies tend to use data that happen to be available. But there is a need to measure OSNs to satisfy specific research questions. How could OSN measurement data be obtained?
- Measurement: focus on fundamentals
 - Time evolution
 - Emergent properties
 - Search for invariants