Big Data + Society = A Data Driven Society?

Danilo Montesi
Department of Computer Science and Engineering
University of Bologna, Italy
danilo.montesi@unibo.it

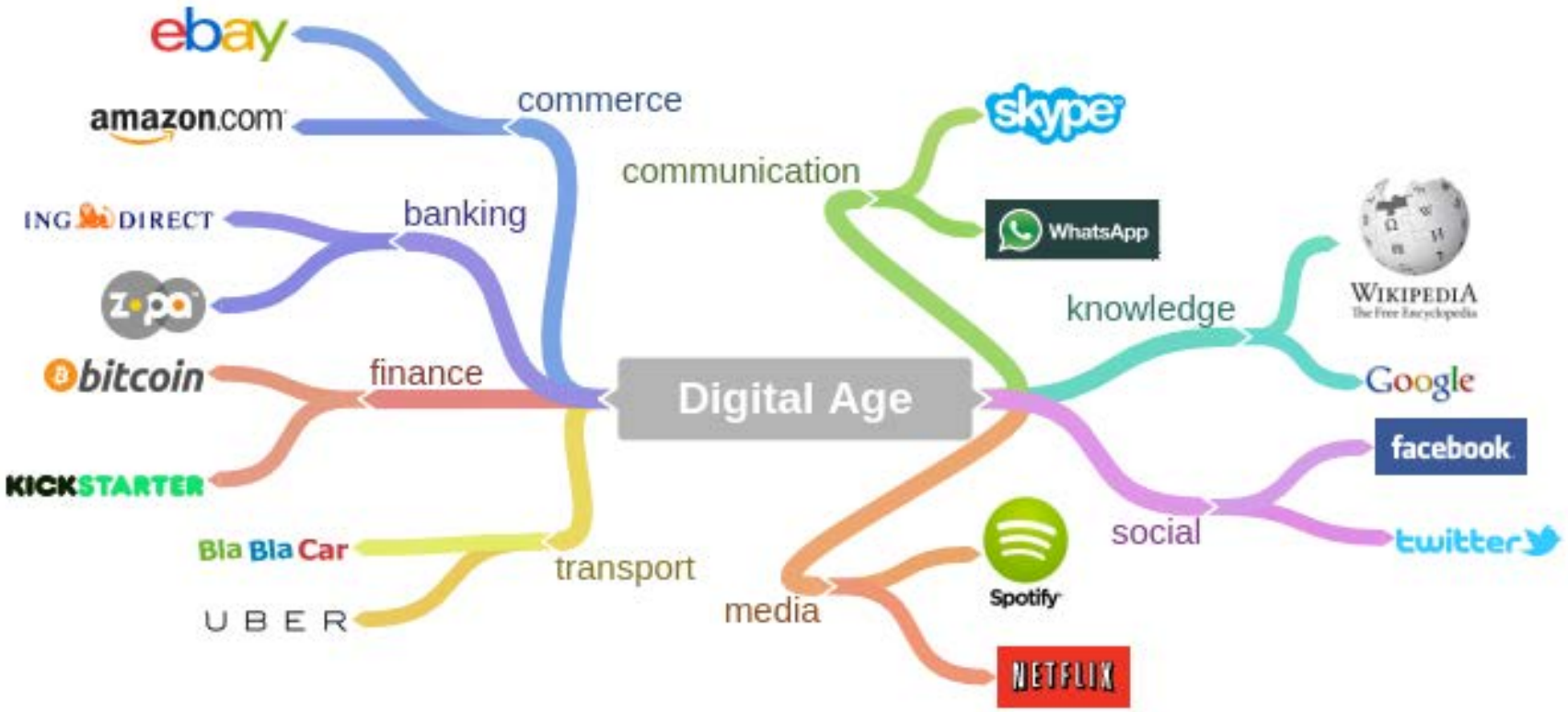
SmartData
University of Bologna

joint contribution with: Stefano Giovanni Rizzo, Flavio Bertini, Rajesh Sharma and Tommaso Ognibene

Outline of the talk

- (Big) Data as production factor
- Searching Data
- Social Networks
- Peer-To-Peer, new business models and sharing economy
- Implications for governments, companies, citizens, et al
- Conclusions

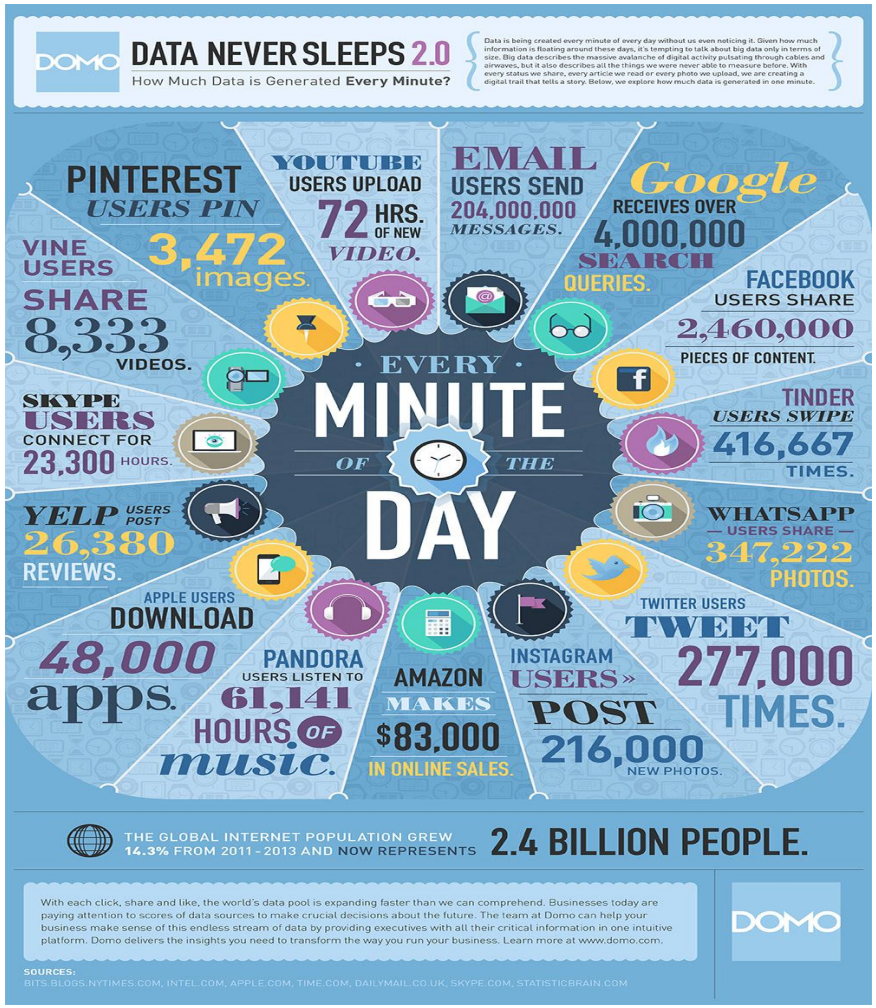
Digital Age



- **Fact**: increasingly fast process of digitalization of any activities
- **Reason:** all the unparalleled benefits of digital computing: precision, velocity, ubiquity, scalability, capillarity, time saving, cost saving and many others
- **Consequence:** growing amount of (digital) data concerning citizens in (almost) every aspects of their lives

How much data? pro-sumer

Pro-ducer



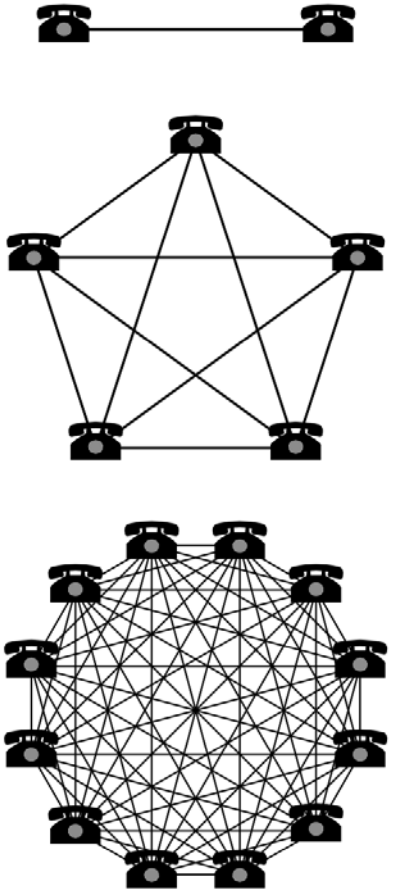
[Source: DOMO]

Con-**sumer**

- On average an American consumes [1]:
  - 5 hours of TV per day
  - 0.73 hours of phone calls
  - 2.86 hours on computers
  - 18.54 GB from computers
- Total consumption/day [1]
  - 11 hours
  - 34 GB of data

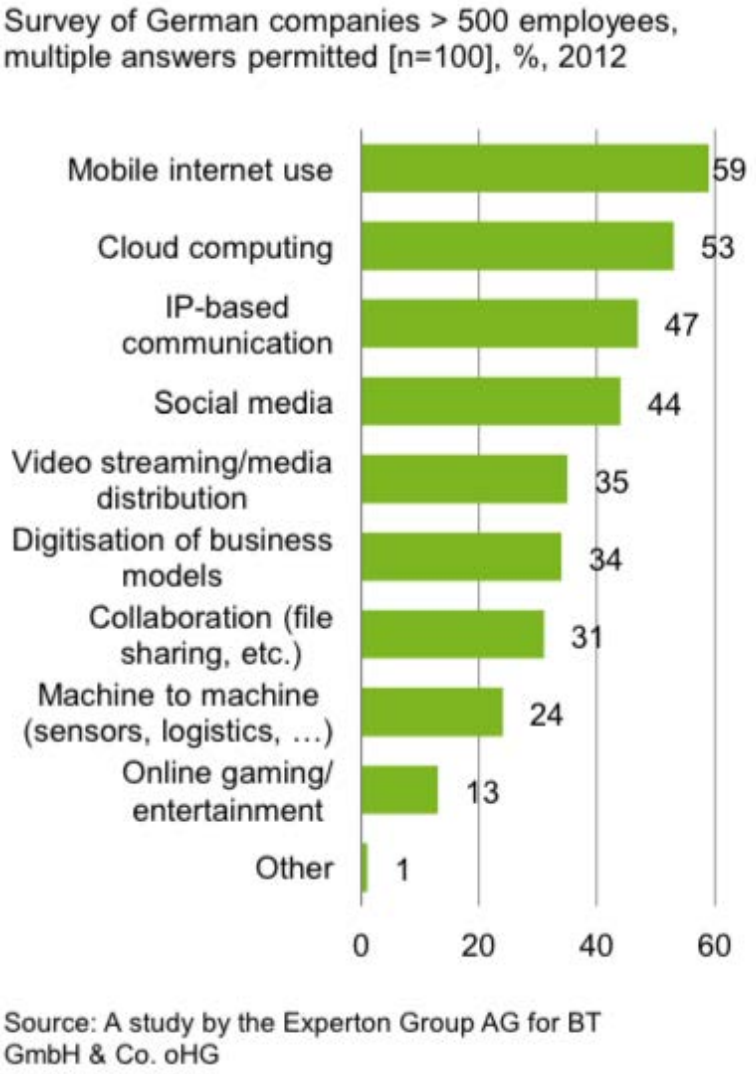
Metcalfe's law (network effect)

- The **value of a network (based business) is proportional to the square of the number of connected users**
- The power of Metcalfe's law:
  - 1 user is worthless
  - 1,000 users are worthless
  - 1,000,000 users are good but not great
  - >100,000,000 users make you very rich
- This effect **is typical of many internet companies**



(Big) Data as production factor

Global drivers for Big Data



Big (Data) Questions

- **Who owns and where are the data?**
  - Google (~90% of online searches in Europe, ~900k servers in 16 main locations)
  - Amazon (~50% of America's book market, ~450k servers in 9 main locations)
  - Facebook (~1 billion members, ~200k servers in 2 main locations)
- **What rights and controls over the data?**
  - Ethical and legal issues concerning **personal data protection** [2]
  - The idea/ideal of **Open Data**
- **How data are evaluated and their consequences?**
  - Feedbacks and reputation **influence the life and success of workers** and **companies** ("**data darwinism**"?)

Data everywhere

- What is the frailty index of population over 65 in Bologna? (healthcare)
- How was Barack Obama's re-election campaign? (society)
- How do you get the right markets and customers? (economy)
- Where and when does the next crime occur? (law)
- How can we identify genetic factors in disease? (scientific)
- We **cannot answer** these questions **without gathering and analyzing data**

Data as production factor

- How important are data?
  - It has gained **economical value**, often more than software
  - It is **challenging several physical business models**
  - Data are becoming **the fourth element of production**
- Factors of production:
  - Land
  - Labour
  - Capital
  - **(Big) Data**

Searching Data

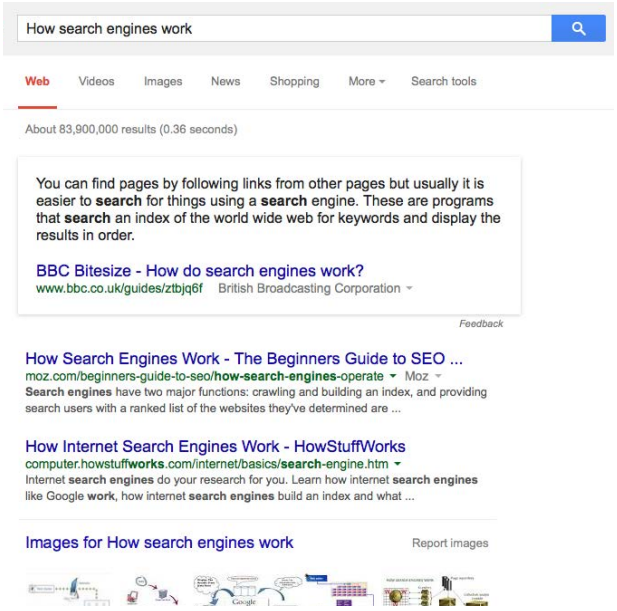
How Search Engines Work

The user inputs a query

The search engine search the query terms in its very large **index**

About 83,900,000 results (0.36 seconds)

Millions of results found are **ranked** accordingly to a notion of **relevance** and are shown as a sorted list

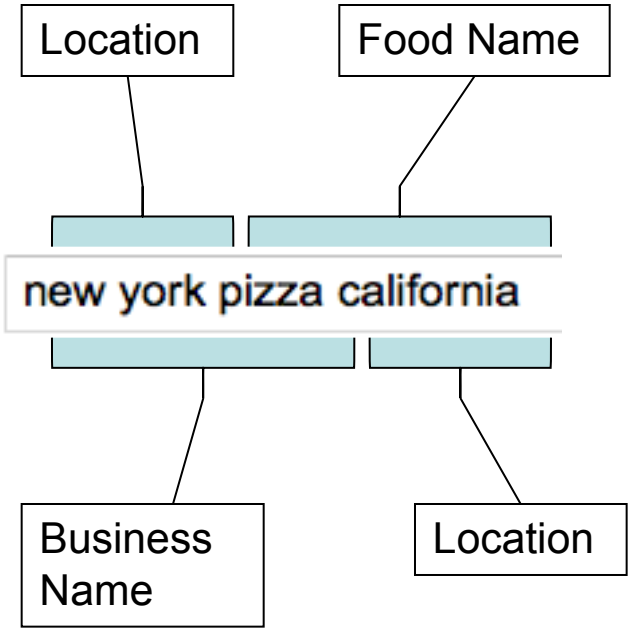


Automated programs (a.k.a. spiders or **crawlers**) scan all pages in the web to build an index

What is Relevant?

- How can a search engine know what the user is looking for?
  - **Information need:** *a desire to locate and obtain information to satisfy a conscious or unconscious need* (Wikipedia)
  - **User query:** the set of consecutive terms formulated by a user to express his information need
  - **Query intent:** the task, goal or intent of a user, expressed by a query. The same query formulated by different users may be the result of different intents

Query intent can be ambiguous:

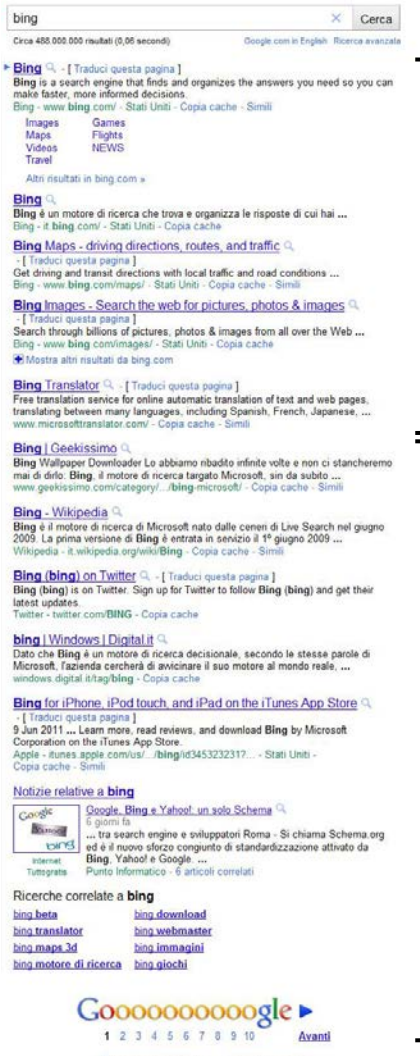


Ranking: Google PageRank

- A relevance notion based solely on term frequencies is not enough to rank billions of documents
- Complex measures are applied to evaluate the quality, reliability and authority of web pages:
  - Measures based **on topological network properties** (such as Google PageRank)
  - Measures based **on different field boosting** (Title, subtitle, body have different weight)
  - Measures based **on semantics and time-space properties** (freshness of a page)
- PageRank set up a **rich-get-richer loop**, whereby few sites dominate the top ranks

Search Engine Results Page

- A Search Engine Results Page (SERP) shows by default **10 results**
- Of these results, on average, the first **5 are visible** without scrolling down
- We will take in consideration the **user behavior on the first SERP** of search engines (the first 10 results)

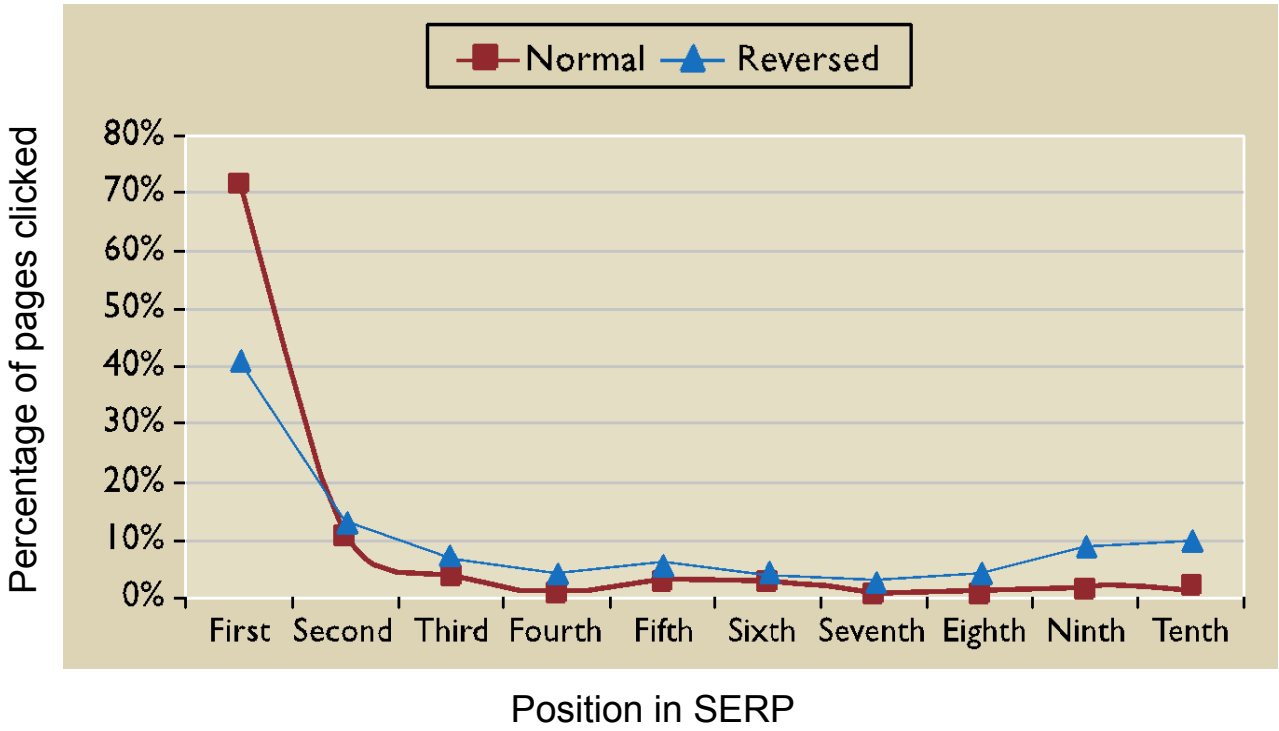


Visible Area

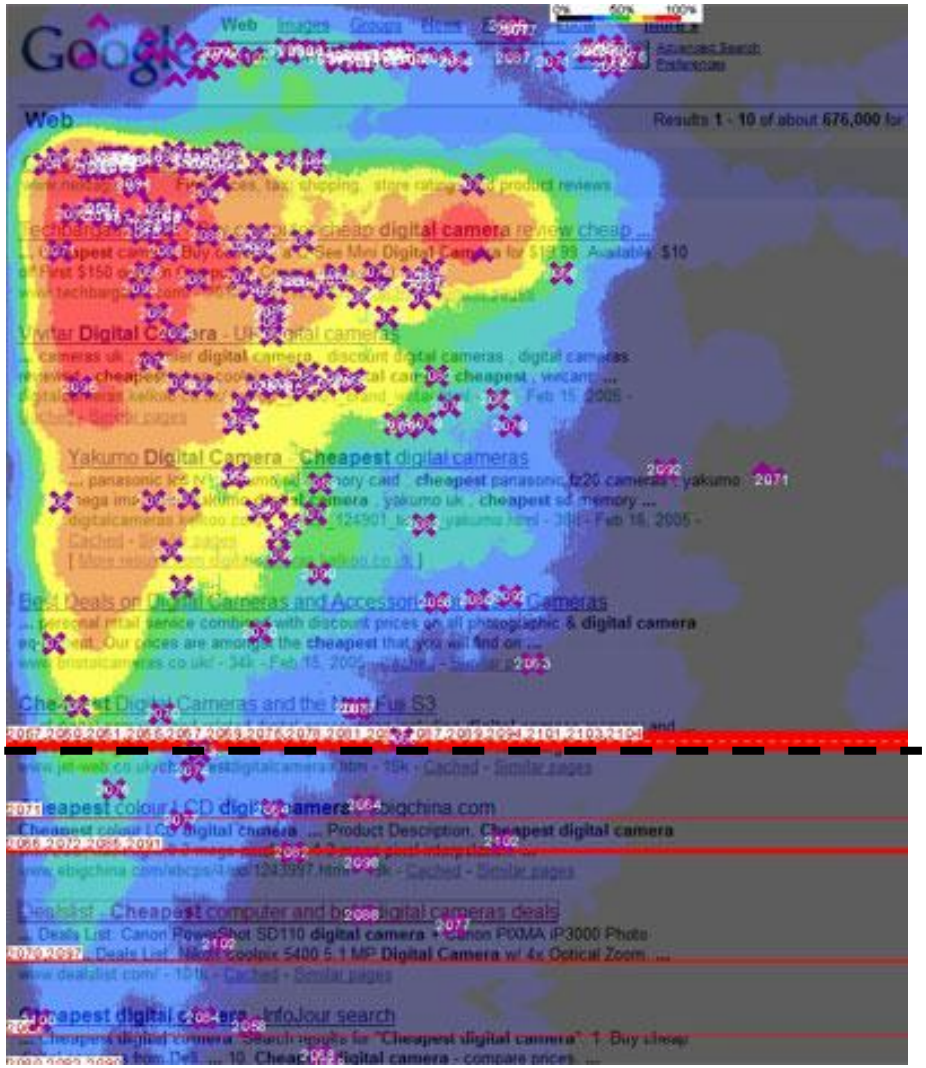
Scroll Area

Result positioning bias

- When searching for information, the average searchers do not judge systematically all the results [4], instead they **simply click on top results**
- Studies have been conducted comparing the users' responses when they received results lists in **normal ordering versus a reversed order**



Eye-tracking: Google *Golden Triangle*

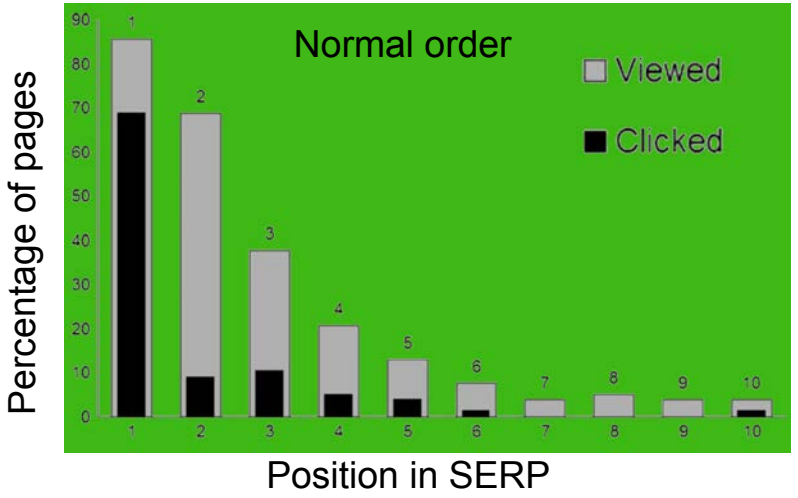


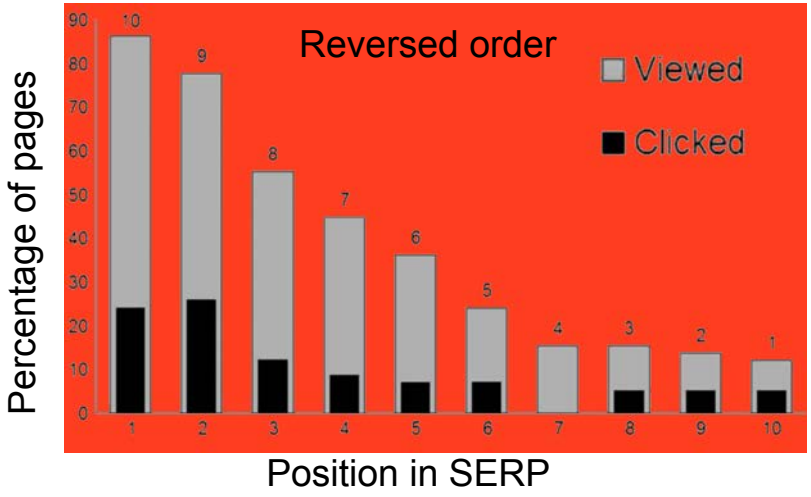
Eye tracking heatmap legend [5]

- **Color**: 0% 50% 100% Percentage of time spent looking the area
- Purple **x** represents a mouse click on the page
- Dotted line **- - -** represents where the page breaks on the computer screen (Visible area/Scroll area)
- Red lines ▬ indicate how far down the page scrolled before leaving the page

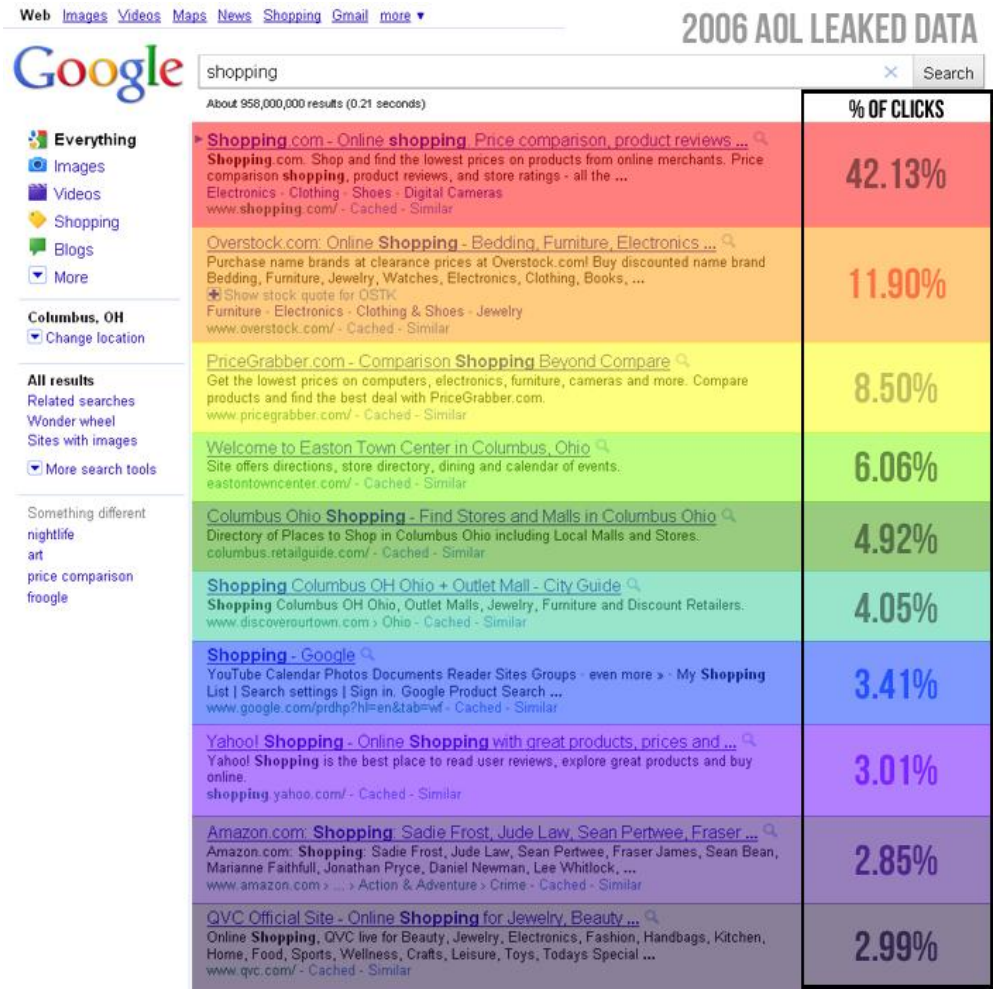
Click and Eye-tracking biasy

- View percentage and mouse clicks are compared in normal vs. reversed order of the first 10 results
- Even though results show some user awareness, **excessive trust in ranking** algorithms [6] may negatively affect smaller positioned websites



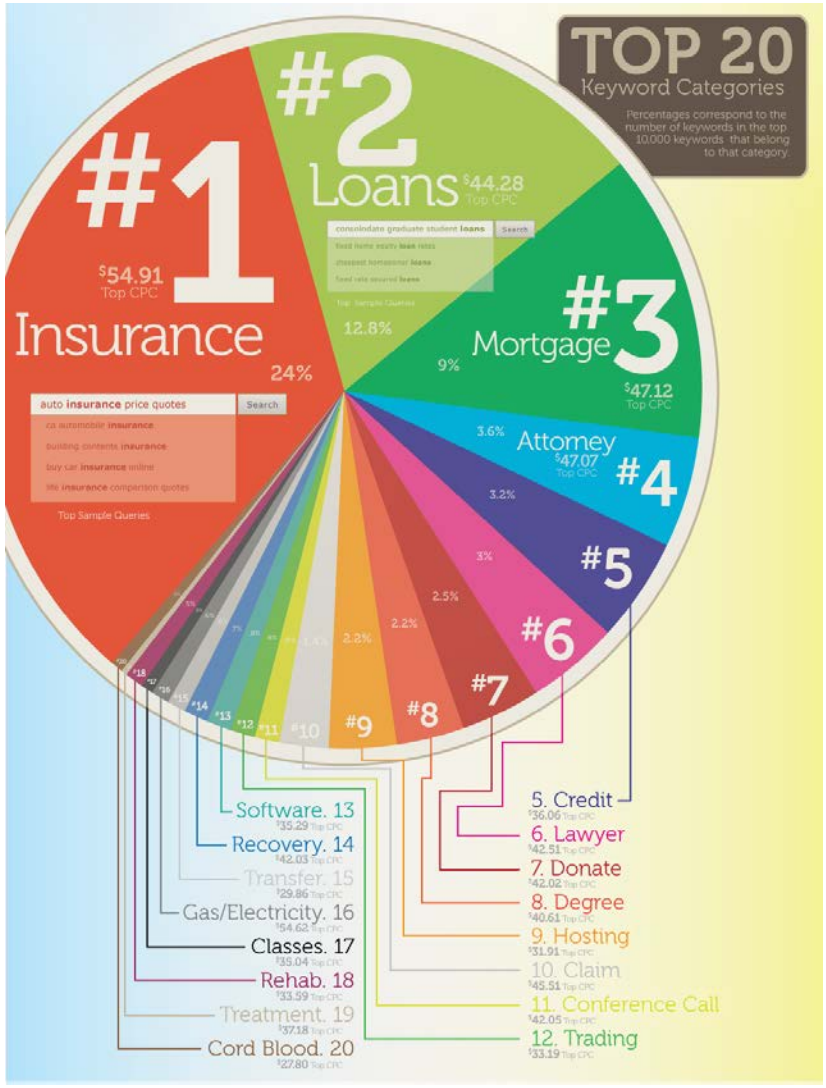


Position impact on business



- Results in 1st position are clicked **more than 10 times** results on 6th position
- Suppose company A and company B are respectively on the 1st and 6th position for a valuable business keyword
- B should buy 10 times the clicks that A gets to obtain the same traffic!
- **How much does a click cost?**

Click economic value



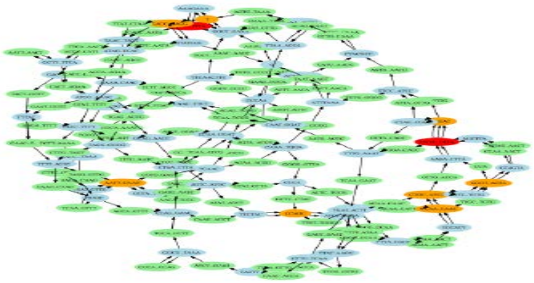
- In web advertising, the cost of a single click, also known as Cost-Per-Click (CPC), is **1$ on average**. However...
- On Google some keywords can be really expensive (Google most expensive keywords 2011):
  - **Insurance: 54.91$**
  - Mortgage: 47.12$
  - Attorney: 47.07
  - Claim: 45.51$
  - Loans: 44.28$
  - Lawyer: 42.51$
- Bing is even more expensive (Bing most expensive keywords 2015):
  - **Lawyers: 109.21$**
  - Attorney: 101.77$
  - Structured settlements: 78.39$

Social Networks

Why Social Networks?

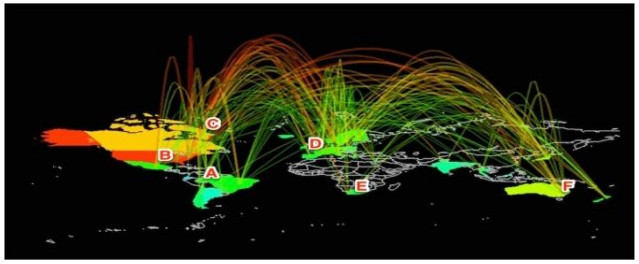
- Social networking is “a gathering of individuals in groups”.
  - Making communities is typical of mankind. *“Man is by nature a political animal” – Aristotle*
  - This tendency to create social connections can be limited by geographical, cultural and time-related obstacles.
- The Web provides an extraordinary and cheap context where very large networks can be created:
  - Users are linked through several kinds of connections, e.g., **friend** (**facebook**), **colleague** (**linkedin**), **follower** (**twitter**).
  - In social networks people can often interact not just with direct connections, but also with the extended network of friends of friends.
  - Messages could include false news or mislead information.

Networks

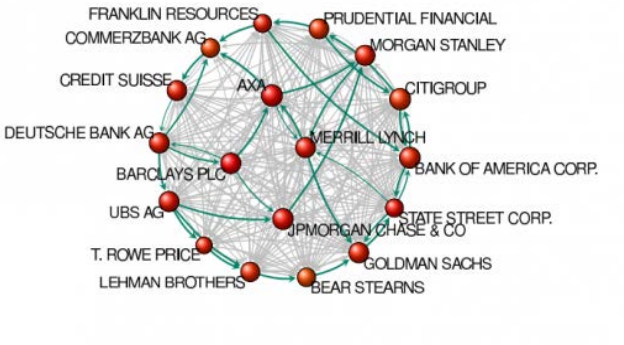


Protein-protein

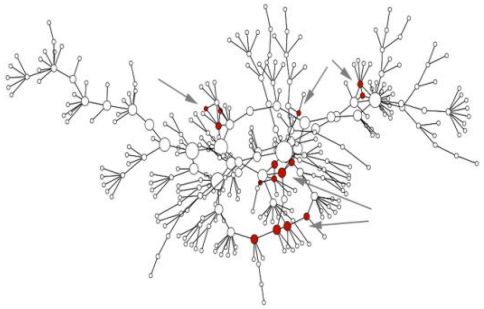
Transportation



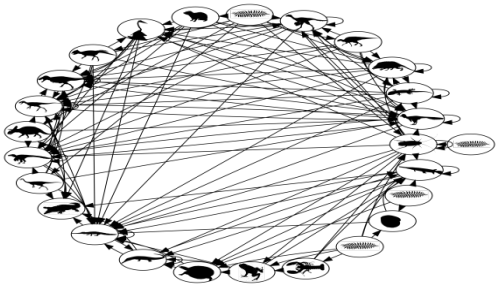
ISP: Router



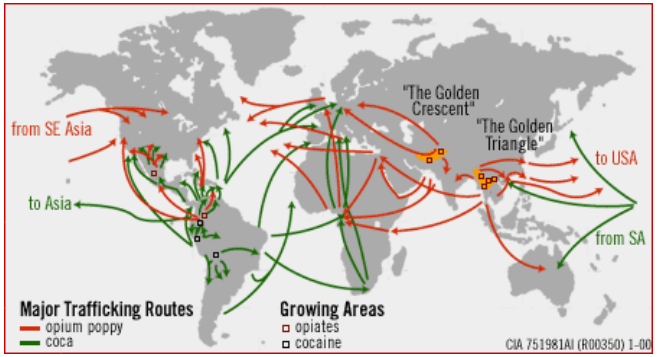
Financial



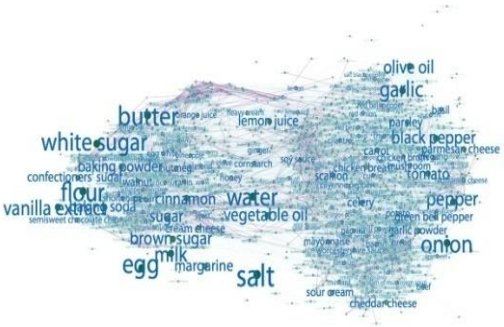
Sexual contact



Food Web



Criminal



Recipe

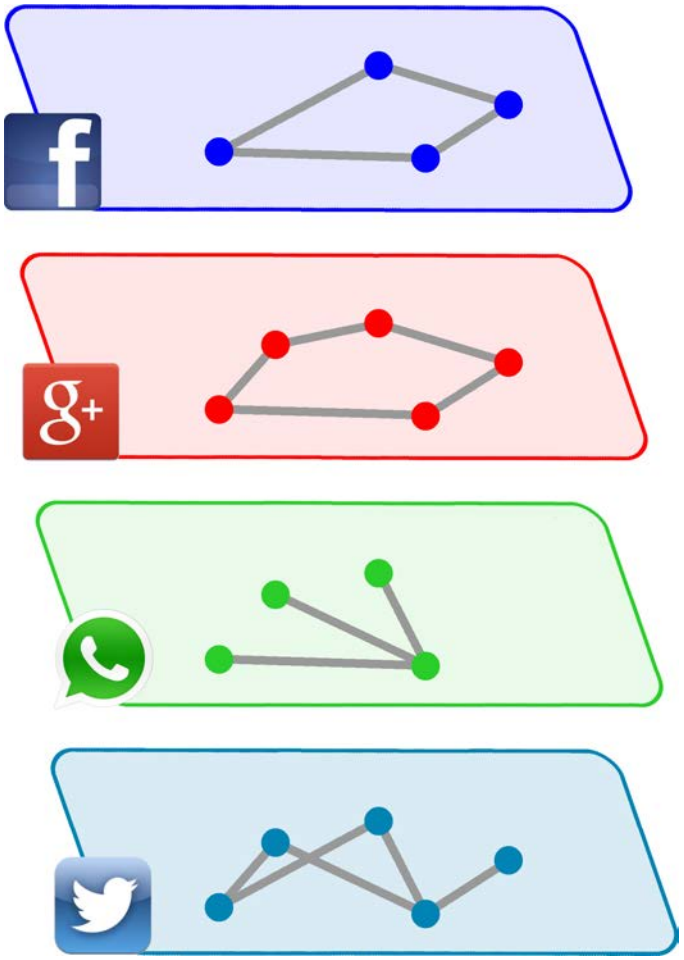
Co-citation

Network Properties

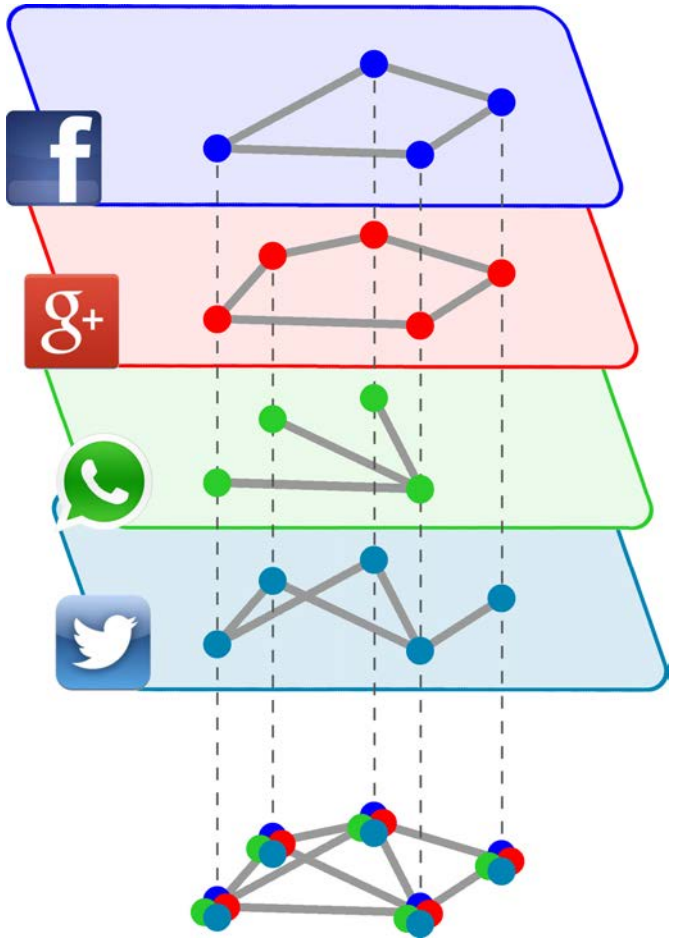
- Cluster coefficient (social triad) [7]
  - useful for **communities detection**
- Degree distribution
  - useful for **hubs & followers identification**
- Average path length
  - useful for **information propagation**
- Diameter
  - useful for defining the **upper bound of the network**
- Centrality measures (betweenness centrality) [8]
  - useful for **detecting the importance of the nodes**

From Networks to Multilayer Networks

Individualistic view



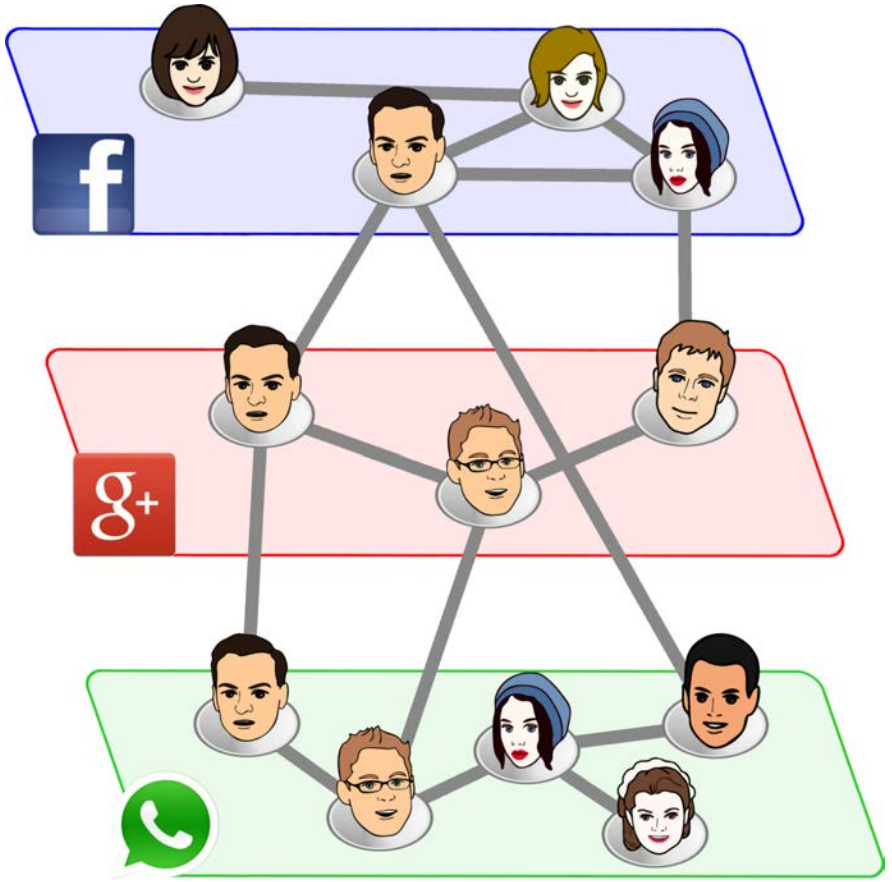
Holistic view [9]



Resultant Network

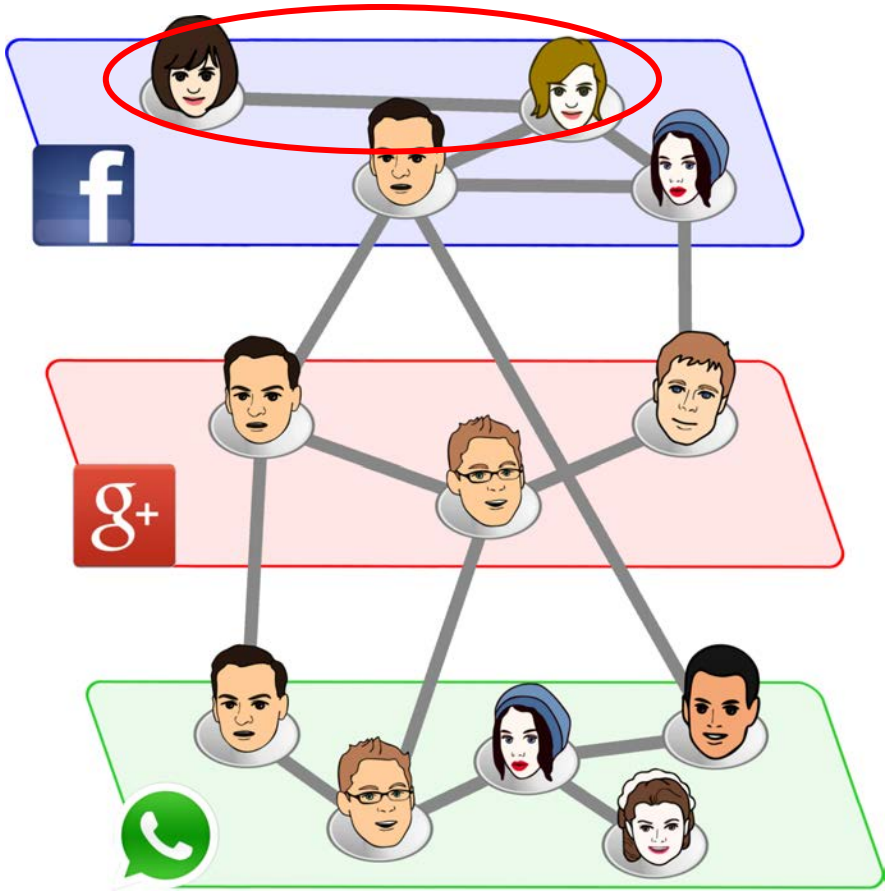
User profiles resolution

- Same user can have different user ids or nickname [10]:
  - same network
  - across networks
- Motivation behind user profile resolution:
  - multilayer networks analysis
  - digital forensics analysis



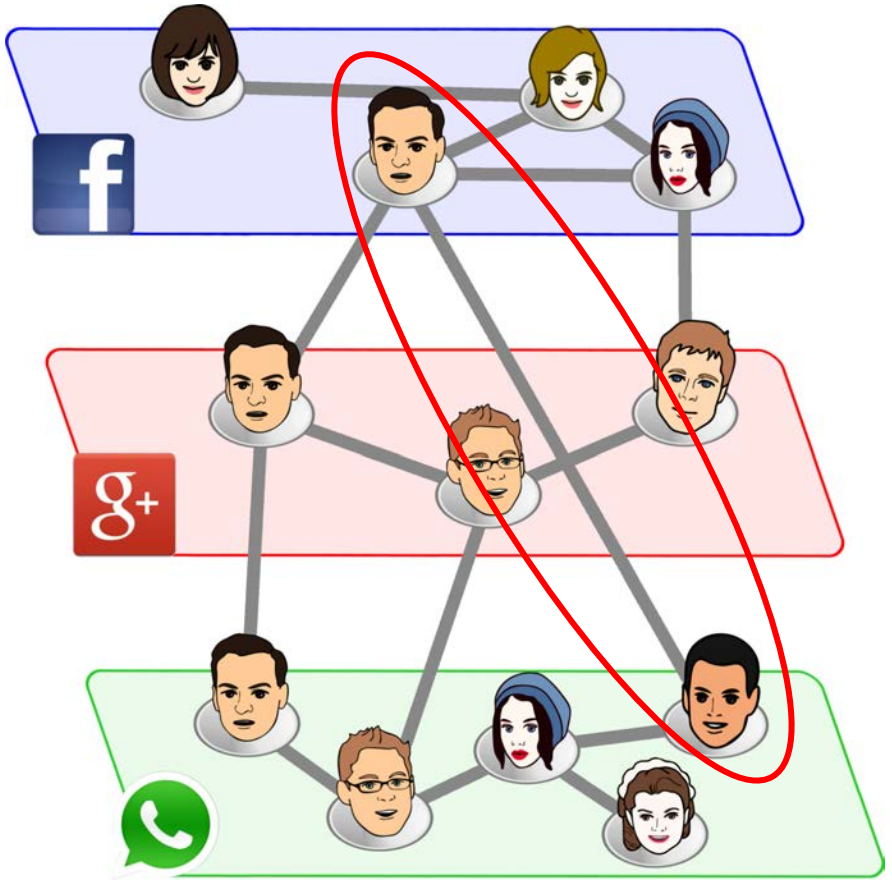
User profiles resolution

- Same user can have different user ids or nickname [10]:
  - same network
  - across networks
- Motivation behind user profile resolution:
  - multilayer networks analysis
  - digital forensics analysis



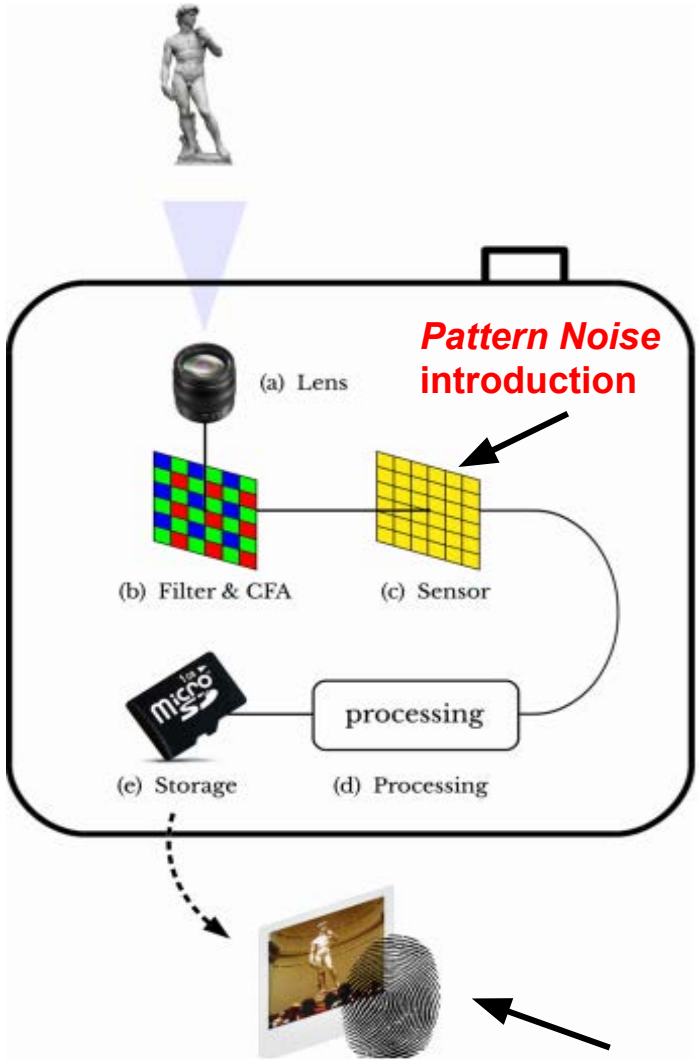
User profiles resolution

- Same user can have different user ids or nickname [10]:
  - same network
  - across networks
- Motivation behind user profile resolution:
  - multilayer networks analysis
  - digital forensics analysis

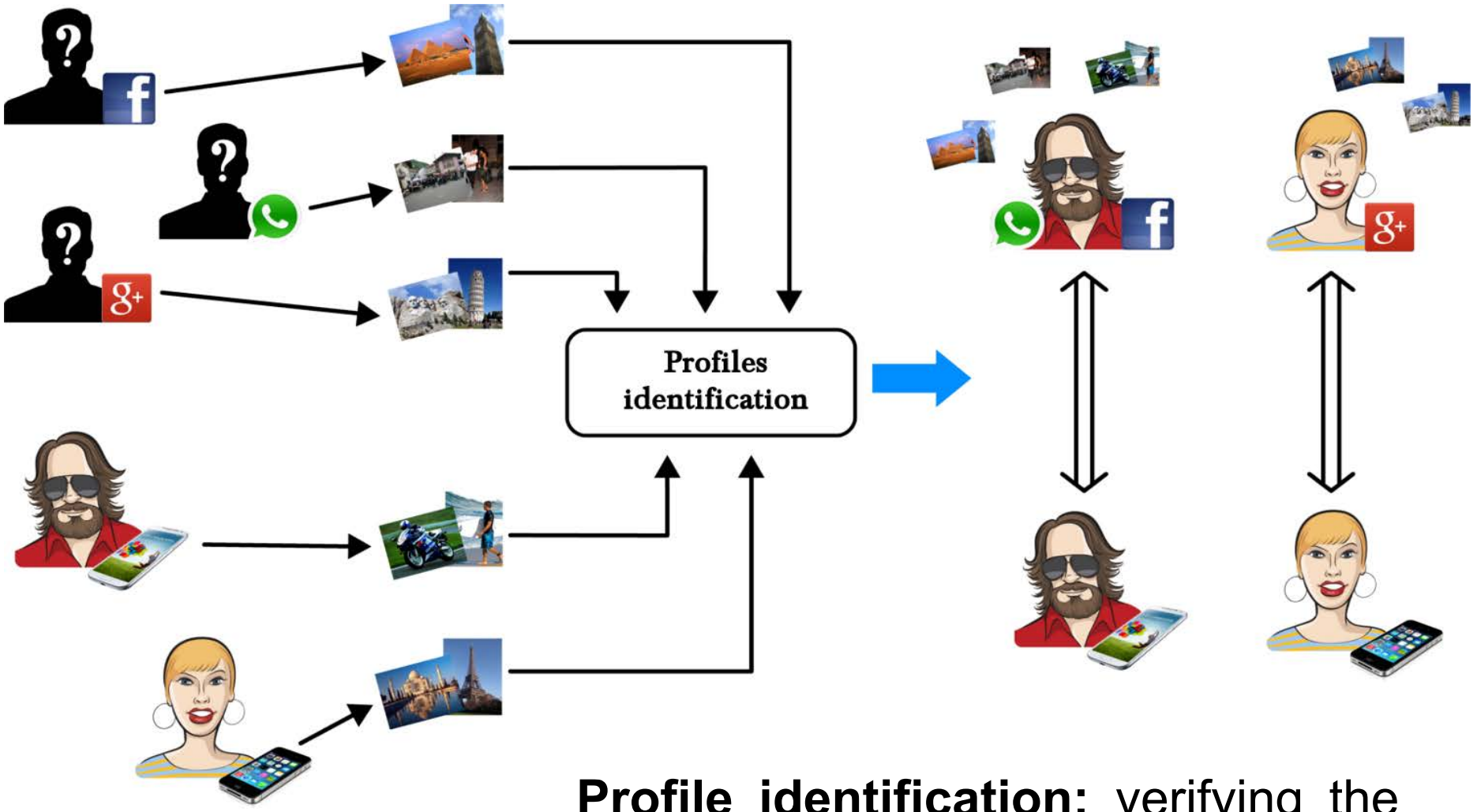


Exploring Smartphone's Camera

- Why the built-in camera?
  - Better than other sensors (e.g. GPS, speakerphone-microphone, etc)
  - Most observed user behavior: pictures sharing
  - Smartphone is more personal than laptop (phone contract hard bound)
- Each image has two components:
  - Signal: the image (what you see)
  - Noise: colour variations not desired
    - Random noise: external factors (e.g. light, humidity, etc)
    - Deterministic noise: *Pattern Noise* useful to create a unique fingerprint without hacking the device

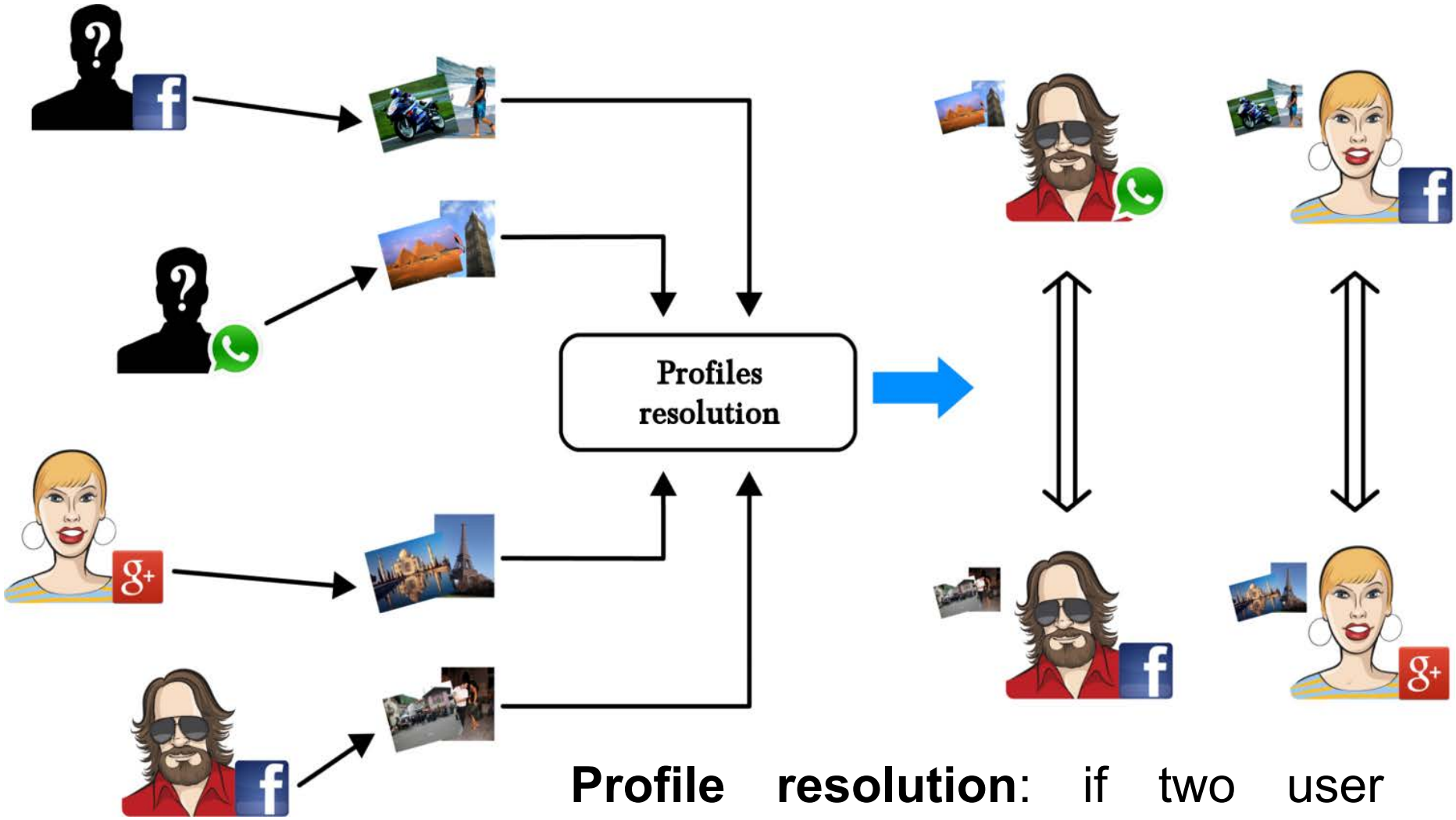


Methodology: profiles identification



Profile identification: verifying the person who is claiming to be.

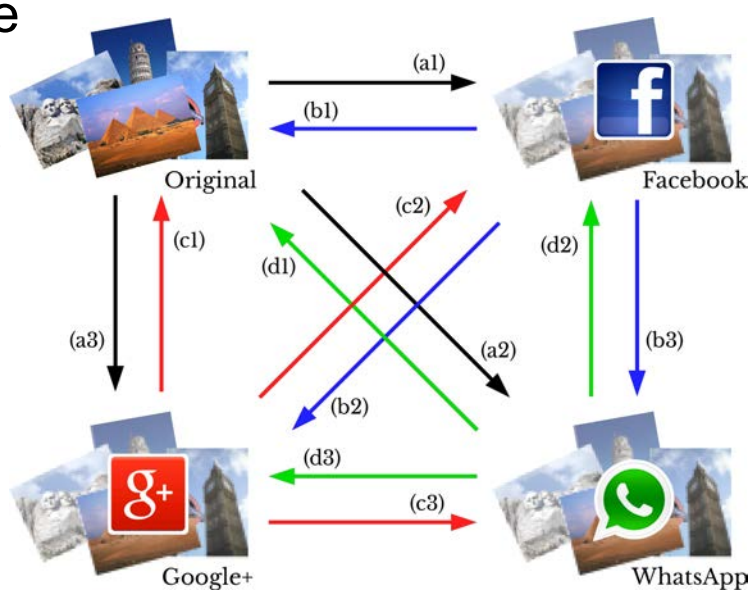
Methodology: profiles resolution



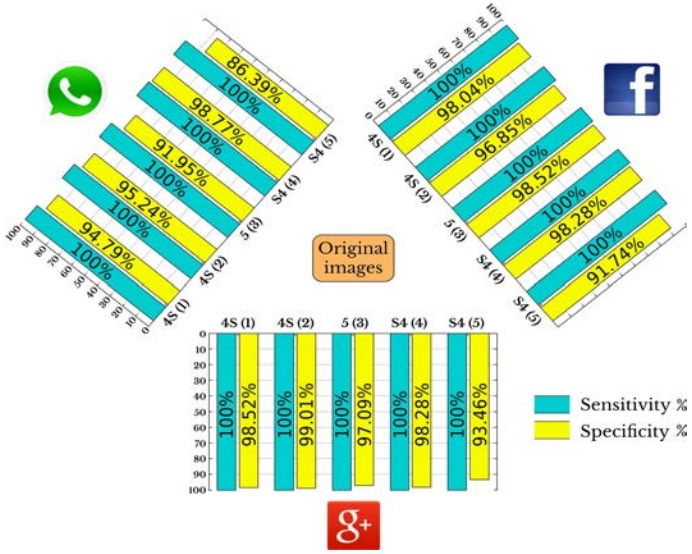
Profile resolution: if two user profiles with different user ids or nicknames belong to the same user.

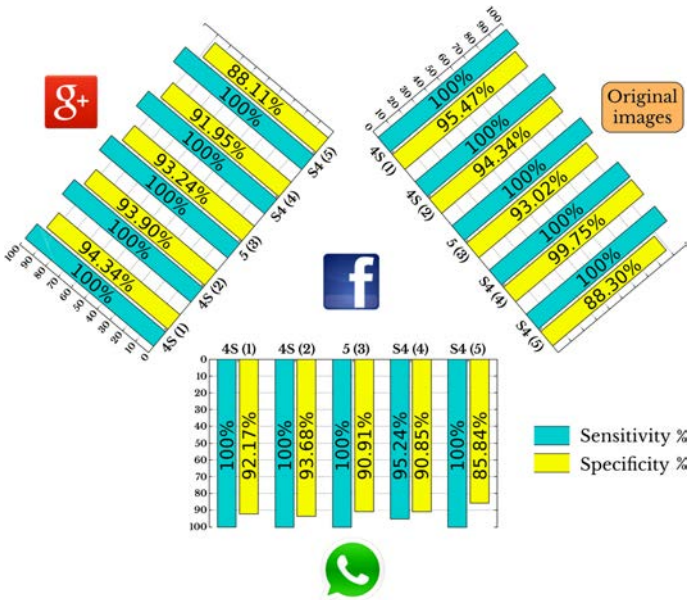
Experimental Setting

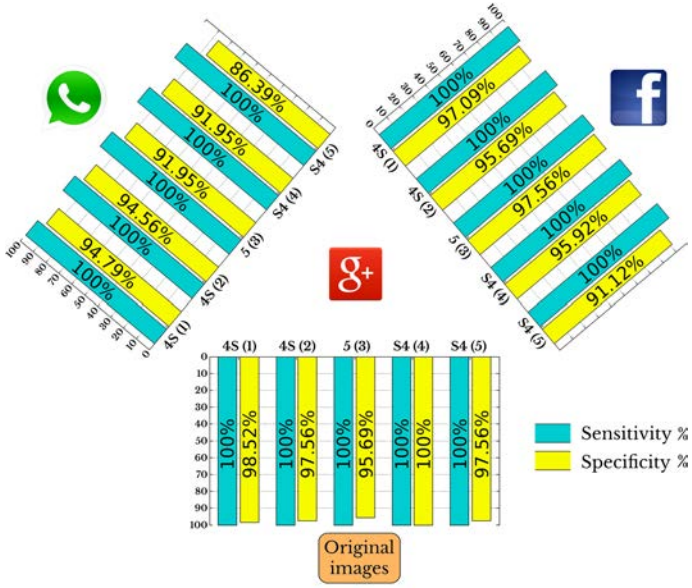
- **Datasets:**
  - 3 different Social Networks: *Facebook*, *Google+*, *Whatsapp*
  - Several smartphones: iPhone 4s, iPhone 5 and Samsung Galaxy S4
  - 1000 images (200 for each smartphone)
- **Combinations tested:**
  - Profiles identification: each Social Network
  - Profiles resolutions: each possible pairing (as shown beside)
- **Measurements:**
  - ***Sensitivity***: the ability to correctly identify right images
  - ***Specificity***: the ability to correctly discard wrong images

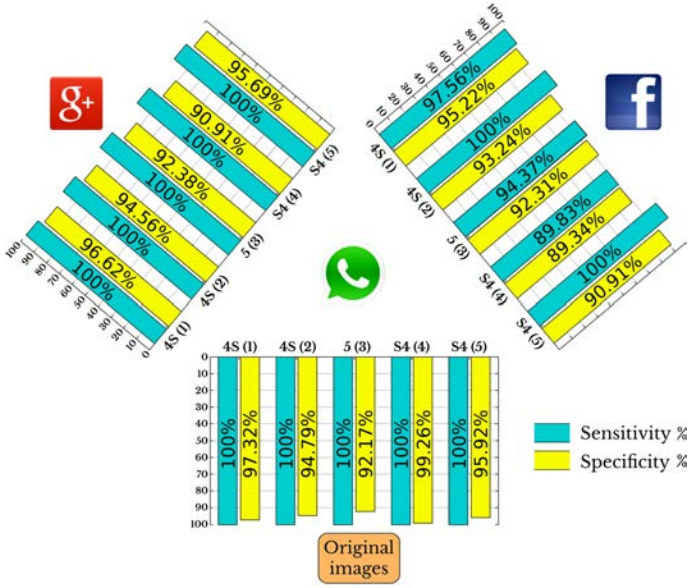


Results

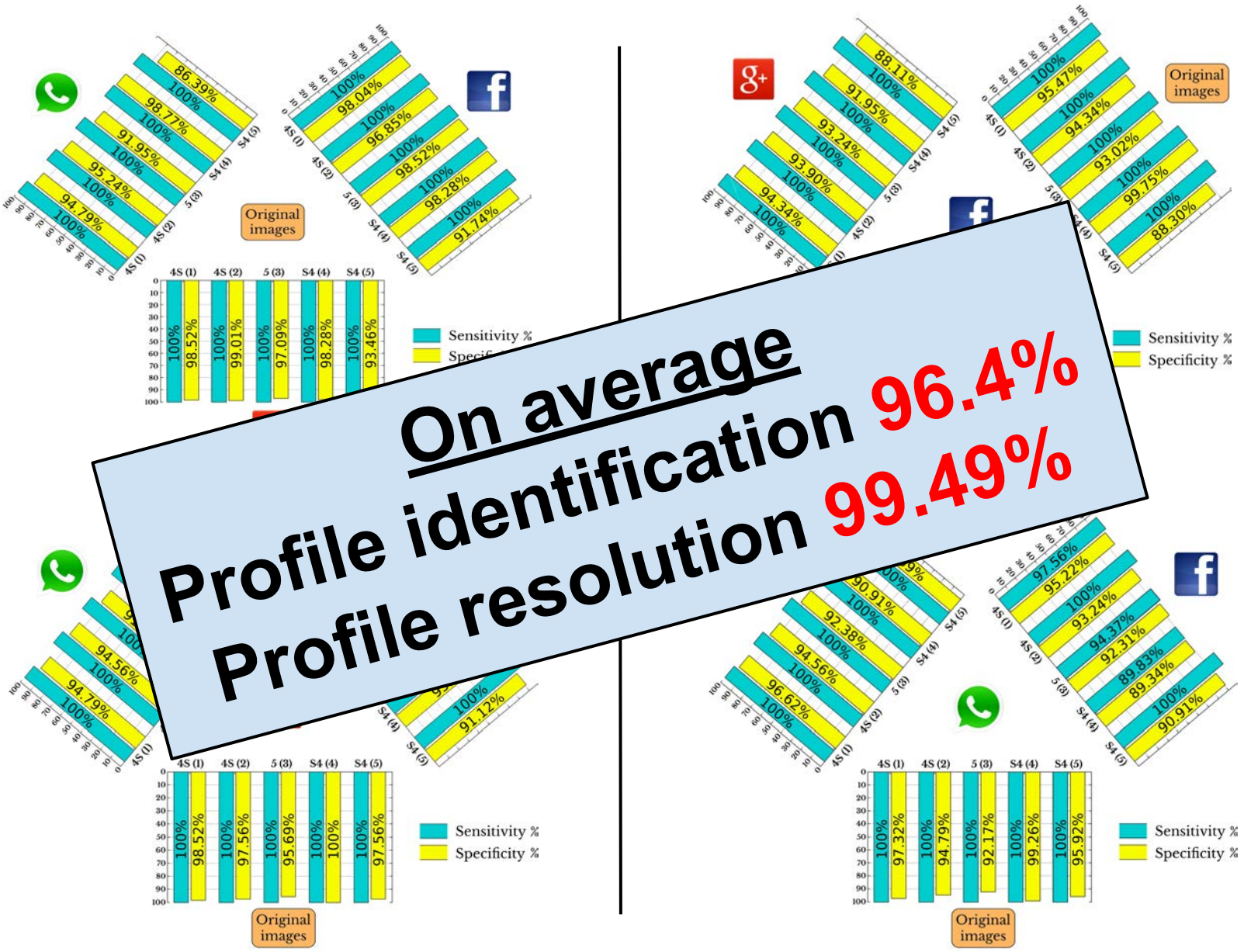

86.39%
100%
98.77%
100%
91.95%
100%
95.24%
100%
94.79%
100%
4S (1)
4S (2)
5 (3)
S4 (4)
S4 (5)
100%
98.04%
100%
96.85%
100%
98.52%
100%
98.28%
100%
91.74%
Original images
4S (1) 4S (2) 5 (3) S4 (4) S4 (5)
100% 98.52%
100% 99.01%
100% 97.09%
100% 98.28%
100% 93.46%
Sensitivity %
Specificity %


88.11%
100%
91.95%
100%
93.24%
100%
93.90%
100%
94.34%
100%
4S (1)
4S (2)
5 (3)
S4 (4)
S4 (5)
100%
95.47%
100%
94.34%
100%
93.02%
100%
99.75%
100%
88.30%
Original images
4S (1) 4S (2) 5 (3) S4 (4) S4 (5)
100% 92.17%
100% 93.68%
100% 90.91%
95.24% 90.85%
100% 85.84%
Sensitivity %
Specificity %


86.39%
100%
91.95%
100%
91.95%
100%
94.56%
100%
94.79%
100%
4S (1)
4S (2)
5 (3)
S4 (4)
S4 (5)
100%
97.09%
100%
95.69%
100%
97.56%
100%
95.92%
100%
91.12%
4S (1) 4S (2) 5 (3) S4 (4) S4 (5)
100% 98.52%
100% 97.56%
100% 95.69%
100% 100%
100% 97.56%
Sensitivity %
Specificity %
Original images


95.69%
100%
90.91%
100%
92.38%
100%
94.56%
100%
96.62%
100%
4S (1)
4S (2)
5 (3)
S4 (4)
S4 (5)
100%
97.56%
95.22%
100%
93.24%
94.37%
92.31%
89.83%
89.34%
100%
90.91%
4S (1) 4S (2) 5 (3) S4 (4) S4 (5)
100% 97.32%
100% 94.79%
100% 92.17%
100% 99.26%
100% 95.92%
Sensitivity %
Specificity %
Original images

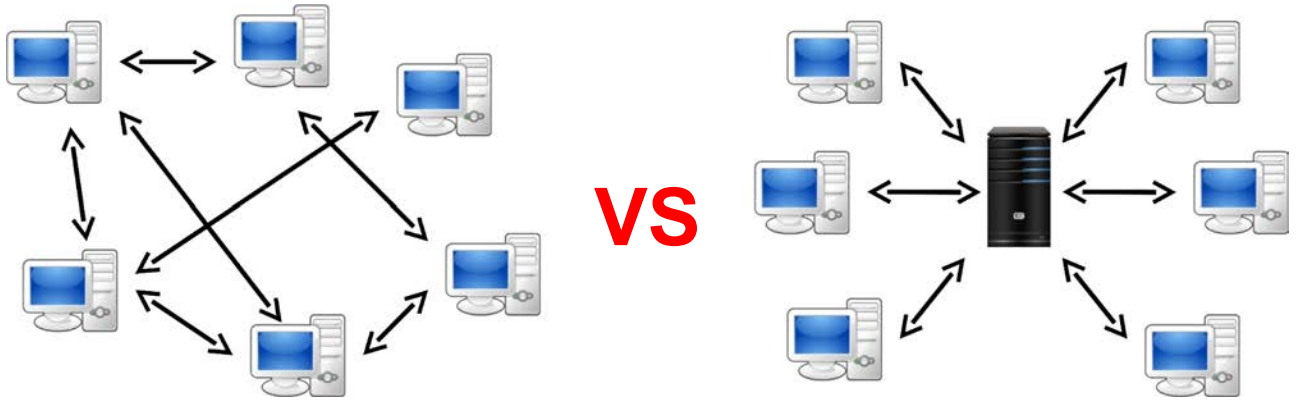
Results


On average
Profile identification 96.4%
Profile resolution 99.49%
Original images
Sensitivity %
Specificity %
4S (1)
4S (2)
5 (3)
S4 (4)
S4 (5)

Peer-To-Peer, new business models and sharing economy

The Peer-To-Peer approach

- The Peer-To-Peer (P2P) is a **network model without a central management unit.** Originally used to share copyrighted material: Napster (1999), eMule (2002) etc
- Each single node has the same functionality of the others.



Peer-To-Peer Client/Server

- P2P has **lower costs and greater flexibility** than the Client/Server network structure.

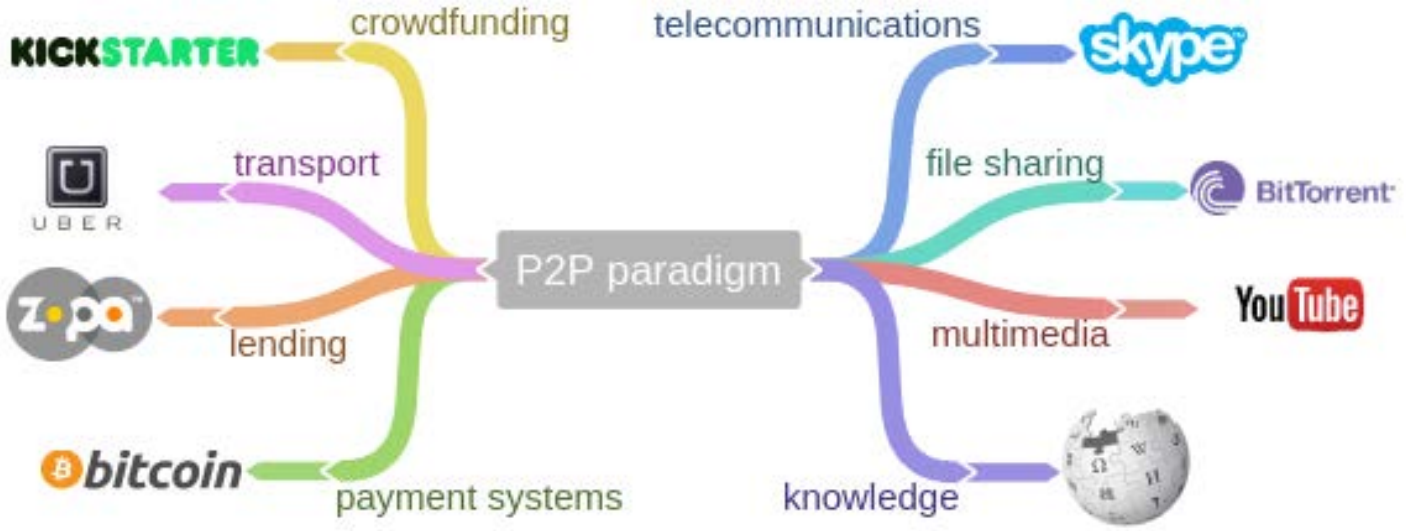
From file sharing to the sharing economy

It's not just a technical paradigm:

- It's a **social paradigm**
- It's an **economical paradigm**

It leads to human disintermediation and the so-called **sharing economy** sometimes **with disruptive effects.**

- (Data) producer meets directly the (data) consumer.



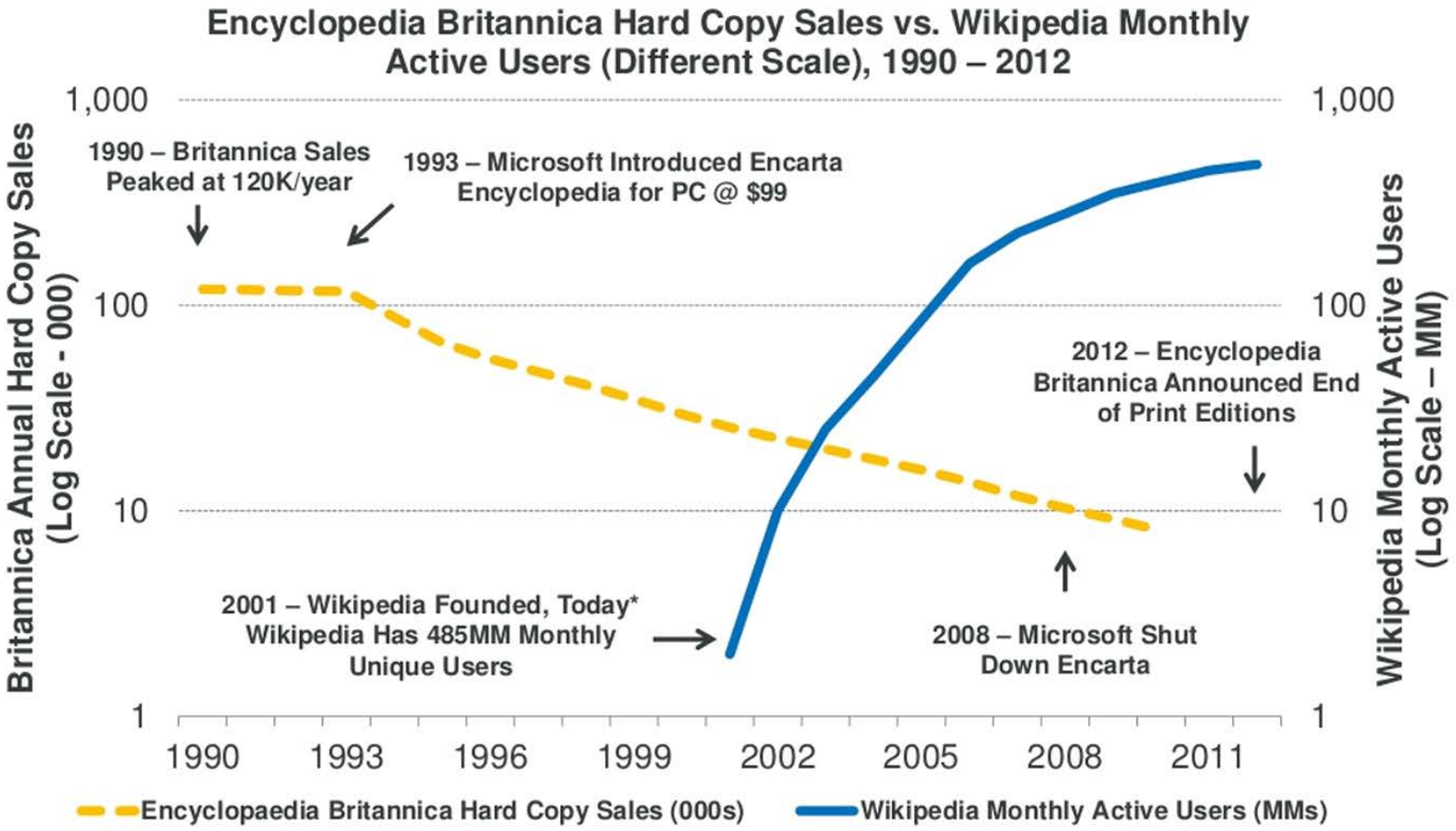
Britannica vs Wikipedia

- The online encyclopedia Wikipedia is the sixth most visited site in the world. The English version has about 4 million of entries, while Britannica has about 230 thousand entries.
- **Uncertainty** about the contents' correctness.
- In 2005, Nature conducted a peer review of scientific entries on Wikipedia and the Encyclopedia Britannica [11]. The study shows that **the result is equivalent**.

	Serious Errors	Inaccuracies/Omissions
Wikipedia	4	162
Britannica	4	123

Disruptive paradigm shift: encyclopedia

244 Years In, Encyclopedia Britannica Went Out of Print in 2012



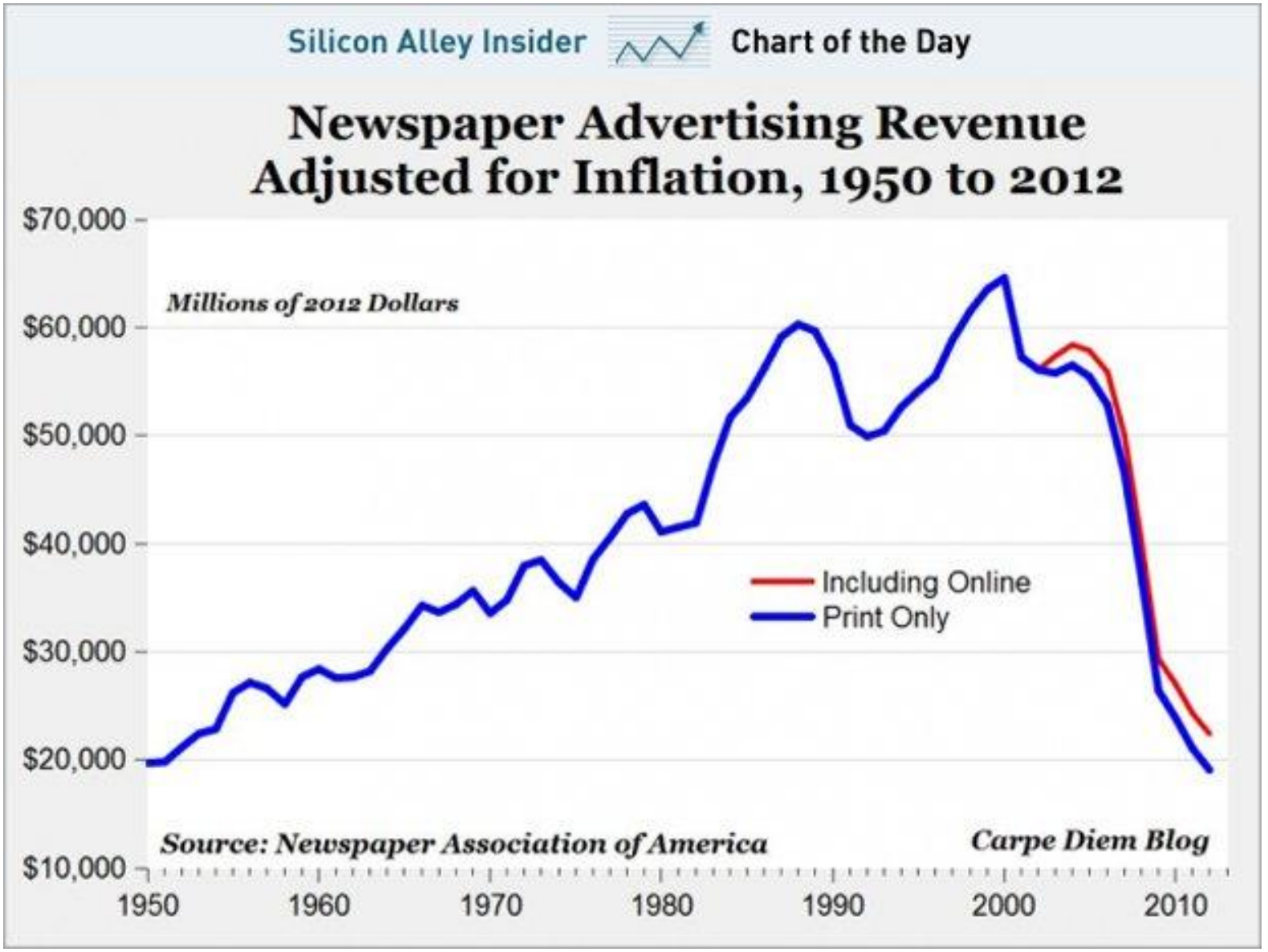
KPCB

Note: *as of 9/12, per comScore global data.
Source: Kellog School of Management, Shane Greenstein and Michelle Devereux, "The Crisis at Encyclopedia Britannica."

Traditional News vs Social News

- The social news are a kind of news aggregators, with a leaner structure than traditional media. The contents come from external websites and blogs. The users rank the news and the most popular ones are highlighted.
- **Uncertainty** about the (missing) editorial board management.
- The Huffington Post is an online news aggregator and blogs (eg. Barack Obama, Madonna, etc) and it is one of the top 100 most visited sites in the world [12].
- In 2013, Reddit had 56 billion visited pages, 731 million unique visitors and 40 million posts [13].

Disruptive paradigm shift: newspaper

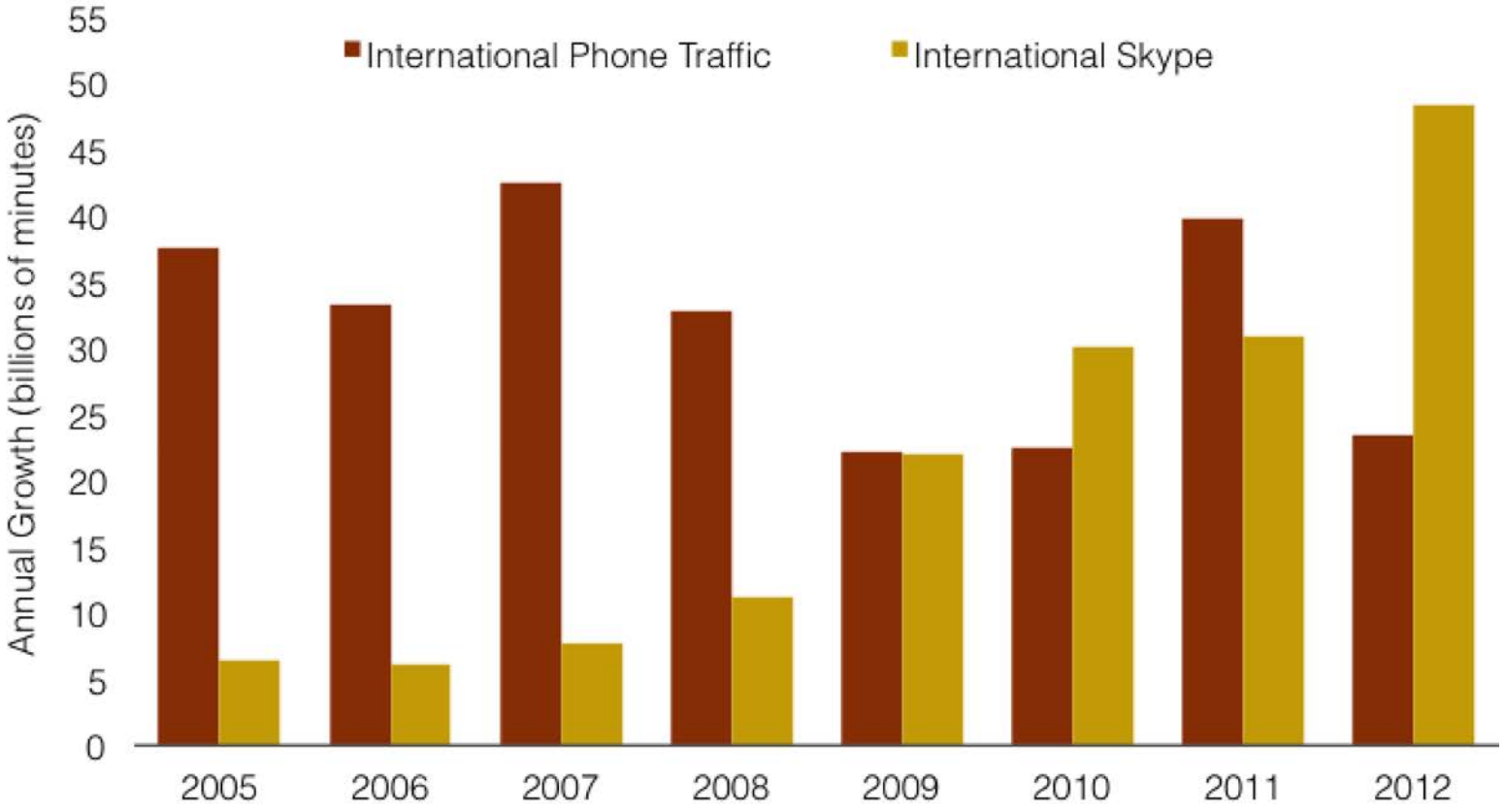


Telco vs Skype

- Skype is a software based on P2P architecture. Users' directory is distributed among the nodes of the network.
- **Uncertainty** about the service quality: disconnections and voice distortion are present, but less and less frequent.
- Conceived in 2002 and the first beta version was released in August 2003. In 2005 it introduced video calls and Skype is currently available on PC, smartphone, smart TV, etc.
- Skype has 650 million of registered users recording **35% of the international calls**. It is **the first operator in the world for international calls.**

Disruptive paradigm shift: call traffic

Increase in International Phone and Skype Traffic, 2005-2012



Notes: ILD traffic reflects TDM and VoIP. Skype traffic growth reflects Skype-to-Skype traffic, including video calls. Skype calls to the PSTN are excluded.

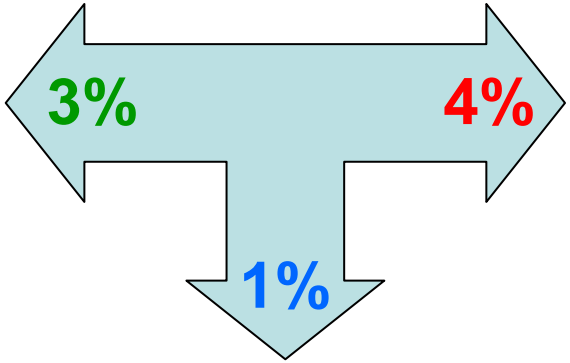
Source: TeleGeography

(Online) Banking vs P2P lending

Bank of America, Barclays, HSBC, citibank

- Zopa is a P2P banking system that allows everyone to disburse and receive loans to each other without intermediary bank. The risk is reduced as the loan is divided on debtors.
- **Uncertainty** on loan repayments.
- Lenders and borrowers have the ability to search and provide the amount and interest rate that fit their needs.

The **lender** decides how much to invest, the rate and the risk level.



The **borrower** is graded by risk and receives the loan by auction or similar.

An annual fee and a (low) percentage of the interest ensure the operation of the **infrastructure**.

Taxi vs Uber

UBER

- Uber supports a transportation network using P2P.
- Allows consumers to submit a trip request, which is routed to crowd-sourced taxi drivers.
- Anyone can be a driver.
- No Taxi driver insurance/licence is needed nor registration with a local authority.
- Uber takes a commission from the driver's earnings.
- **Uncertainty** about qualified and safe drivers.
- Founded in 2009, in 2015, Uber shows off 1 million of daily trips and 50'000 monthly new drivers. It is estimated that Uber will generate 10 billion dollars in revenue by the end of 2015 [14].

£/€/$ vs Bitcoin

- The Bitcoin is a digital currency based on P2P without a central managing authority
- Users are not identified by their name but by a "Bitcoin Address". Thus, the exchanges are anonymous
- **Uncertainty** about the missing of a central entity and consequent sharp price fluctuations and limited supply
- Bitcoin was designed in 2008
- In March 2014, the total capitalization is about 8 billion dollars, with about 70,000 transactions per day
- The average time to confirm a transaction is <10 min

Implications for governments, companies and citizens

Implication: the crisis of the State

- There is a **critical conflict** between the concepts of State and Internet [16]:
  - The concept of State is founded by the principle of **national sovereignty**
  - The concept of the Internet is guided by the principles of **trans-nationality** and **decentralization**
- This conflict weakens the effectiveness of legal regulations and jurisprudence showing the **limit of national jurisdiction**
- A solution would be a strong and possibly unanimous **international regulations and jurisprudence**

Implication: concentration of power

- Driven by the **network effect**, several markets are experiencing the emerging of a handful of dominant companies, leading to serious **implications on the labor market**. For instance:
  - Thousands of newspapers threatened by Google News
  - Thousands of bookshops threatened by Amazon
  - Thousands of taxi drivers threatened by Uber
- Concentration of power also leads to **social control**: relying on certain technologies, **these technologies can influence our life**:
  - I don't need to remember how to drive from A to B; I just need to know how to use Google Maps.

Implication: liquid information economy

- **Competition between legal and fiscal systems:**
  - In a liquid economy, a company can **easily move** its production everywhere, accordingly to its best interests (tax, labor, environmental laws), challenging the effectiveness of domestic law
- **Big Data as a competitive advantage:**
  - A company's competitiveness is determined by its ability to **make better decisions with better information**
- **Ownership replaced by access right:**
  - Once a first copy of information has been produced, additional copies cost nothing. Ownership becomes increasingly marginal, while the key to business success is the **access right to information**

Implication: data are a virtual currency

- Example scenarios:
  - Millions of users make billions of online searches with Google, **freely**
  - Millions of users share billions of messages with Facebook, **freely**
  - Millions of users listen to music for billions of hours with Spotify, **freely**
- Why freely? Because, if you're not paying for the product, **the product is you**:
  - Information retrieved by Google is not totally objective. **The order in which results are displayed has a price**
  - Facebook and Spotify make money mainly through **target advertising**: they show you things you may want to buy because they know what you may want to buy
- They collect, analyse and sell personal data to various companies. Thus, **personal data are a new virtual currency**

Implication: the future of employment

- There are growing concerns regarding the so-called **technological unemployment** [15]
- There were similar concerns also in the early days of the Industrial Revolution, with respect to the **automation of production**
- But there are **two main differences**:
  - The digitalization of labor is evolving at an **unprecedented pace**
  - Digitalization carries the ability to **work from anywhere in the world**. Companies can choose where to put their **workplace** accordingly to their interests

Bring on the personal trainers

Probability that computerisation will lead to job losses within the next two decades, 2013 (1=certain)

Job	Probability
Recreational therapists	0.003
Dentists	0.004
Athletic trainers	0.007
Clergy	0.008
Chemical engineers	0.02
Editors	0.06
Firefighters	0.17
Actors	0.37
Health technologists	0.40
Economists	0.43
Commercial pilots	0.55
Machinists	0.65
Word processors and typists	0.81
Real estate sales agents	0.86
Technical writers	0.89
Retail salespersons	0.92
Accountants and auditors	0.94
Telemarketers	0.99

Implication: a data driven life

- Consider the implications of unveiling and mapping worldwide the information concerning **life expectancy**, **wealth, crime rate**, **employment rate**, **effectiveness of public services**, **cost of living** and so on.
  - Would you **buy an house** in a neighborhood with an **higher than average crime rate**?
  - Would you **choose an hospital** with a **higher than average number of lawsuits for medical malpractice**?
  - Would you send your **kids to a school** with a **lower than average rating**?

Conclusions

- An **unprecedented** volume of data are used
- Data are becoming the **fourth factor of production**
- Digital age is producing the **winner takes it all effect** with deep implications on governaments, enterprise, employment and citizens
- Open problems are: **privacy, competition, control, sharing**
- **Personal data are already a virtual currency**
- With enough personal data **you can create comprehensive profiles of people**

Thank you! - Questions?

Danilo Montesi
Department of Computer Science and Engineering
University of Bologna, Italy
danilo.montesi@unibo.it

SmartData
University of Bologna

joint contribution with: Stefano Giovanni Rizzo, Flavio Bertini, Rajesh Sharma and Tommaso Ognibene

References

[1] R. E. Bohn and J. E. Short, *How Much Information?* Report on America Consumers, Global Information Industry Center, University of California, 2009.

[2] K. Davis, *Ethics of Big Data: Balancing Risk and Innovation.* O'Reilly, 2012.

[3] P. Lawrence, B. Sergey, M. Rajeev and W. Terry, *The PageRank Citation Ranking: Bringing Order to the Web.* Unique to ILPubs Technical Report. Stanford InfoLab, 1999.

[4] K. M. T. O'Brien and M. Smyth, *Are people biased in their use of search engines?* Communication ACM 51(2), 2008.

[5] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini and B. Pan, *Eye tracking and online search: Lessons learned and challenges ahead.* Journal of the American Society for Information Science and Technology 59(7), 2008.

[6] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay and L. Granka, *In Google We Trust: Users' Decisions on Rank, Position, and Relevance.* Journal of Computer-Mediated Communication 12(3), 2007.

[7] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks.* Nature vol. 393 n° 6684, 1998.

[8] L.C. Freeman, *A set of measures of centrality based on betweenness.* Sociometry 40, 1977.

[9] M. Salehi, R. Sharma, M. Marzolla, M. Magnani, P. Siyari and D. Montesi, *Spreading processes in Multilayer Networks.* IEEE Transactions on Network Science and Engineering, 2015, to appear.

[10] F. Bertini, A. Ianni, R. Sharma and D. Montesi, *Matching user profiles in multilayer networks through smartphone camera fingerprint.* XXXV Sunbelt Conference of the International Network for Social Network Analysis (INSNA), 2015, to appear.

[11] J. Giles, *Internet encyclopaedias go head to head: Jimmy Wales' Wikipedia comes close to Britannica in terms of the accuracy of its science entries.* Nature vol. 438 n° 7070, 2005.

[12] Alexa Traffic Ranks, *How popular is huffingtonpost.com?* http://www.alexa.com/siteinfo/huffingtonpost.com

[13] C. Smith, *By the numbers: 40+ amazing Reddit statistics.* Digital Marketing Stats, 2015.

[14] A. Shontell, *Uber Is Generating A Staggering Amount Of Revenue.* Business Insider, 2014.

[15] C. B. Frey and M. A. Osborne, *The future of employment: how susceptible are jobs to computerisation?*, 2013.

[16] T. Owen, *Disruptive Power. The Crisis of the State in the Digital Age.* Oxford Studies in Digital Politics, 2015.