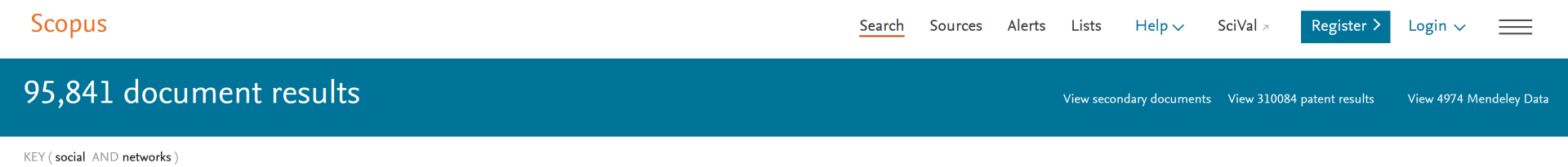
Knowledge Discovery in the social network era: An overview of (big) data analytics approaches for social networks

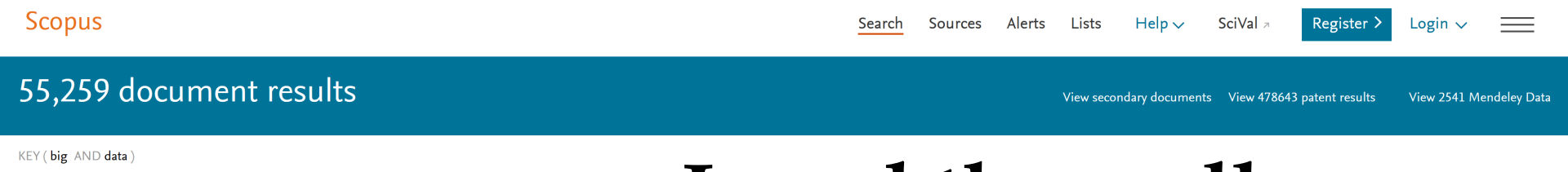
Elio Masciari

Big Data 2019, Los Angeles
December 09-12, 2019

Before starting…

- Social Networks and Big Data related topics are generating a lot of research effort





I read them all…

I swear ☺

Outline

- A brief history of Social Networks (SN)
- The Big Data Challenges
- Social Networks (SN)

  Big Data Features
- What happened so far
- Conclusions

How humans became social network addicted…

- **2,800,000 BC**: Humans first appear on the earth. Somehow, they manage to learn to express themselves and communicate with each other despite a complete lack of funny memes and emojis except raw graffiti
- **550 BC**: The world's first postal service is created in Assyria. The phrase "the cheque's in the mail" is coined
- **1792**: The telegraph is invented. The first telegraph message ever sent? "New telegraph, who's this?"

(Source: Phrasee)

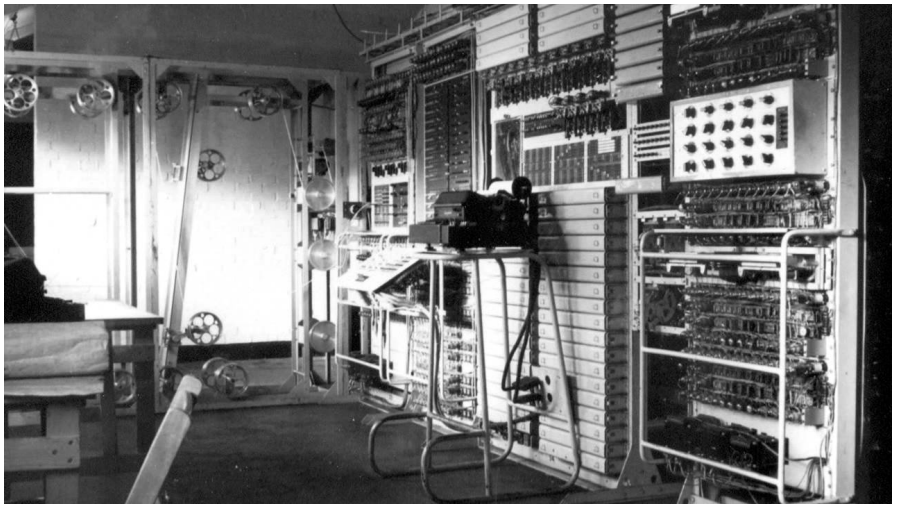
The first “real” step!

- 1839: The world’s first “selfie” is taken by amateur chemist and photography enthusiast Robert Cornelius

Then…

- 1890: The telephone is invented. Sadly, it is unable to take photographs or send text messages yet, rendering it almost completely useless…
- 1891: The radio, an early precursor to Spotify, is invented

Some years later…



- **1940s**: The world's first "supercomputers" are built. The world's scientists begin developing ways for those supercomputers to communicate with each other. Their dream? To finally be able to publicly share photos of their lunch!!!

Let's get in touch…

- **1960s**: The earliest forms of the "internet" begin to appear, but are extremely limited in both scale and scope.
- **1969 – August 5th, 1991**: ("The good old days") There is no worldwide web. Human communication and social networking remain analog, and friends continue to pester the world's citizens with stacks of photos from their vacations.

Getting ready to socialize…

- **August 6th, 1991**: The worldwide web is unleashed upon an unsuspecting public. Meanwhile, a 4-year-old "Star Wars Kid" is already in training for his internet debut.
- **1997**: The world's first social networking site: "Six Degrees" is born. Remember it? Few really do…
- **January 2000**: The millennium bug shuts down everything and ushers in a second dark age. Just kidding. Nothing happened.
- **2001**: Six Degrees shuts down forever. Suprisingly enough because they tried to target university students…Wait, who?

Things go faster and faster…

- **December 2002**: Linkedin is born. The world's headhunters and hiring agencies lick their chops in anticipation of how easy their jobs are about to become.
- **March 2002**: Friendster is launched. People with very few friends suddenly have lots of friends. The socially awkward of the world rejoice.
- **July 2002**: Friendster reaches 3 million users. The world's adults shake their heads in dismay
- **May 2003**: Image sharing site Photobucket is launched. Millions of useless digital photos of people's cats suddenly have a purpose

And faster…

myspace®
a place for friends

flickr™

- **August 2003** Myspace is founded, and the awesome power of the social network is slowly revealed to the world.
- **February 2004**: Facebook goes live. The world has no idea what's about to take place.
- **February 2004**: Photo sharing site Flickr is launched. Digital camera sales increase exponentially…
- **April 2004**: Facebook Ads are launched to support exponential growth.

And faster…

- **June 2005**: The world's angriest website, Reddit, is launched
- **February 2005**: YouTube is born, sounding the death knell of America's Funniest Home Videos
- **July 2005**: Myspace reaches 22 million users and is growing at a rate of 2 million per month. Everyone can see this bubble will never, ever burst.
- **October 2006**: YouTube is acquired by Google for $1.65 billion, and gets progressively suckier with every passing year.

And faster…

- **March 2006**: Twitter goes live, paving the way for Donald Trump's eventual ascension to the Whitehouse…
- **September 2006**: The Facebook "Newsfeed" goes live. Facebook's grip on what the world's citizens see and hear tightens.
- **August 2007**: The hashtag (#) debuts on Twitter #awesomeidea

And faster…

- **October 2008**: Spotify goes live. The zombie hand of the music industry bursts through the ground in front of its tombstone
- **February 2009**: Facebook introduces the "like" button. Liking something is instantly transformed from a matter of personal taste to a social necessity. The question "why didn't you like my…" is uttered for the first time.
- **May 2009**: Facebook surpasses Myspace in user count for the first time. The writing is now on the wall
- **September 2009**: Facebook announces that it is cashflow positive for the first time. Mark Zuckerberg high-fives himself in the mirror repeatedly for 6 straight days.

And faster…

- **March 2010**: Pinterest goes live. The world's craft enthusiasts rejoice.
- **October 2010**: Instagram is launched and hits 1 million users by December. The selfie takes its first steps toward become the bane of humankind.
- **July 2011**: Snapchat is launched just in time. Everyone's parents now have Facebook accounts, which is awful.
- **April 2012**: Facebook acquires Instagram for $1 billion, almost cornering the global market of duck-faced selfies and butt photos
- **October 2012**: Facebook reaches 1 billion active users. The world wakes up to the fact that a for-profit corporation now owns one of its most important communication channels, and that there's nothing anyone can do about it.

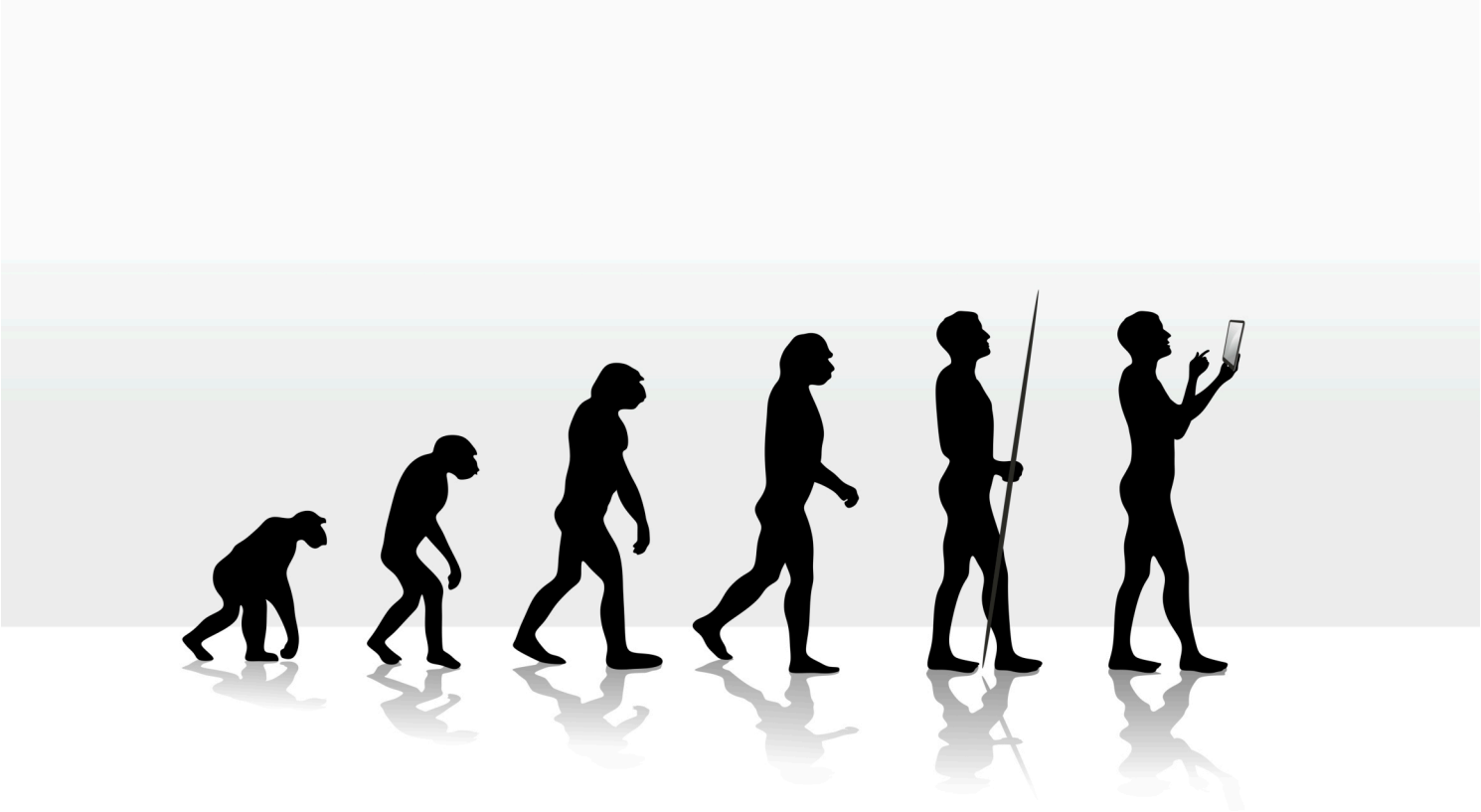
And faster…

- **December 2012**: Twitter hits 140 million active users. It remains unclear how many of them are actually Russian troll-bots yet…
- **February 2013**: Snapchat users are now sending 60 million snaps per day. Sadly, approximately 58.5 million of them are insufferable.
- **January 2013**: Myspace re-launches with a new website, mobile app, and endorsement from Justin Timberlake. Internet users flood back in droves. Just kidding. Nobody cares.
- **March 2014**: Instagram reaches 200 million active users. People are now going to specific restaurants specifically to take photos of interesting-looking foods.
- **April 2015**: Snapchat reaches 100 million active users, largely made up of millennials. The world's parents barely take notice and stick to Facebook.
- **May 2015**: Facebook enables GIFs, which is awesome.
- **June 2015**: Friendster shuts down. Wait, who were they again?

And faster…

- **September 2015**: Snapchat introduces it's "filters" feature, allowing users to add animal ears and rainbow puke to their snaps. The internet LOVES it.
- **June 2016**: Instagram announces it has reached 500 million active users, many of them creeps.
- **February 2016**: Time Inc buys Myspace (for some reason)
- **September 2016**: Snapchat re-brands itself as "Snap inc" and releases smart sunglasses called "Spectacles". Nobody cares.
- **March 2018**: It is revealed that Cambridge Analytica harvested troves of user data without their consent and used this data for political purposes. Facebook stock plummets
- **April 2018**: Mark Zuckerberg testifies before Congress. The world sees how weird becoming one of the world's richest humans can make you.

To summarize in a snapshot …



Nowadays reality: (*Big Social*) Data, Data, Data

- Large volumes, Large diversification of data features referred as **Big Data**:
  - Users
  - Connections
  - Actions
  - Contents
  - Sensors
  - Mobile devices
- (Online) Social networks are the new “Petroleum” that calls for more and more advanced analysis strategies

Nowadays reality: (Big Social) Data, Data, Data

- Online Social Networking indeed is a novel paradigm within the Big Data Analytics framework, where massive amounts of information about heterogeneous social facts and interactions are stored, at a very high variability rate. Important problems such as Influence Diffusion and Maximization, Community Detection, User Recommendation require skills from both research fields making the research activity quite challenging

Source: Big data in Social networks (Picariello 2018)

Outline

- A brief history of Social Networks (SN)
- **The Big Data Challenges**
- Social Networks (SN) Big Data Features
- What happened so far
- Some systems
- Conclusions

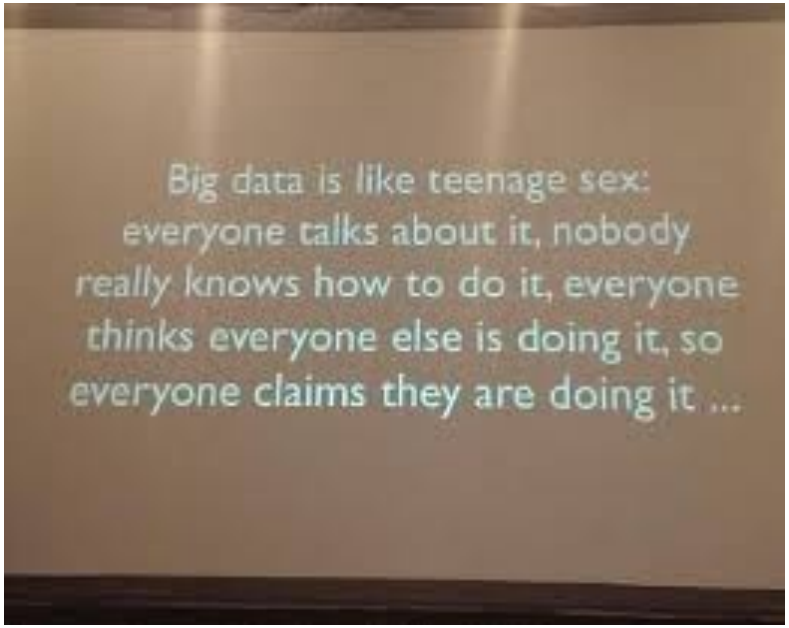
I know what you are thinking…

Big Data (in the mass culture)

Big Data (a pessimistic vision)

- Large volumes, Large diversification, High Speed:
  - 3V initial paradigm
    - Volume
    - Velocity
    - Variety

Big Data (an optimistic vision)



- Add more V:
  - Veracity
  - Variability

Big Data (for real life)

"Let's say you want to save millions of dollars — you just push this button here..."

The last V:

- **Value**

Outline

- A brief history of Social Networks (SN)
- The Big Data Challenges
- **Social Networks (SN) Big Data Features**
- What happened so far
- Conclusions

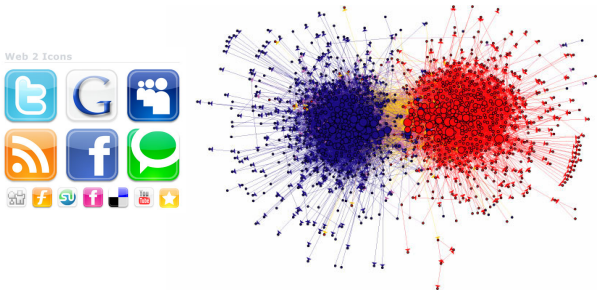
Why social networks are key Big Data providers?

- They continuously generate an enormous quantity of heterogeneous data gathering the most valuable information: user habits!
- Do ut Des strategy of the big companies like Amazon, Apple, Facebook, Google, Microsoft

Which (Big) Data?

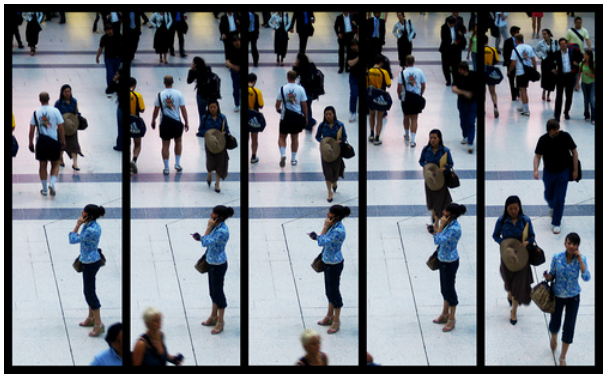
Shopping patterns & lifestyle

Relationships & social ties



Desires, opinions, sentiments

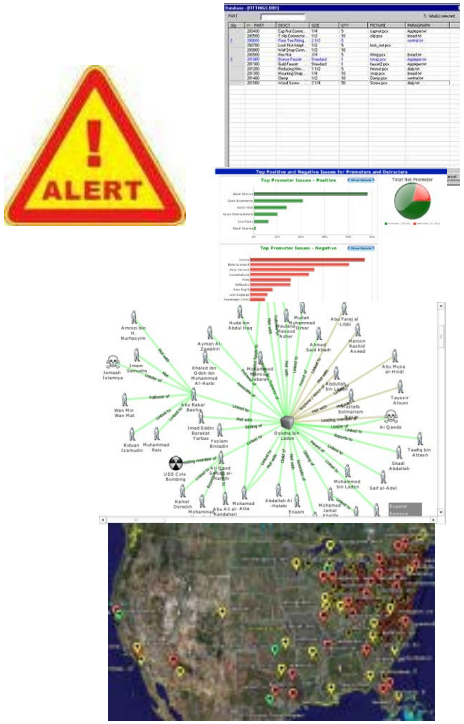
Movements



Big challenges

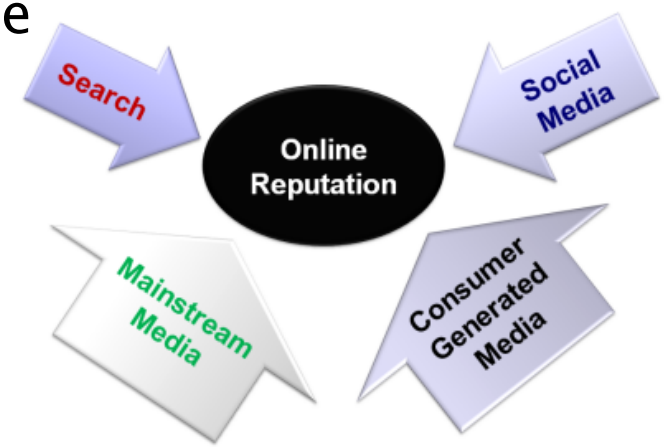
Consumer Generated,
Not Edited,
Not Authenticated

Actionable Intelligence



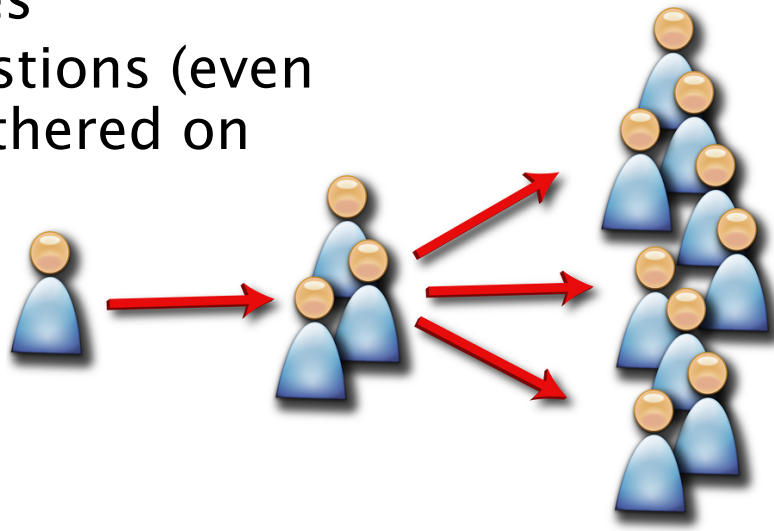
What industries search for…

- Consumer Brand Analytics
  - What is the opinion on my brand?
- Marketing Communications
  - Measure the effectiveness of marketing campaigns
- Product reviews
  - Measure what people say
    - Easy to use, comfortable, adequate price, …



An example

- Viral marketing:
  - Personalized recommendations
- The role of online forums:
  - 79.2% of forum members help other users to make decisions on product purchases
  - 65% of forum members share suggestions (even offline) based on the information gathered on forums



http://www.socialmediaexaminer.com/new-studies-show-value-of-social-me

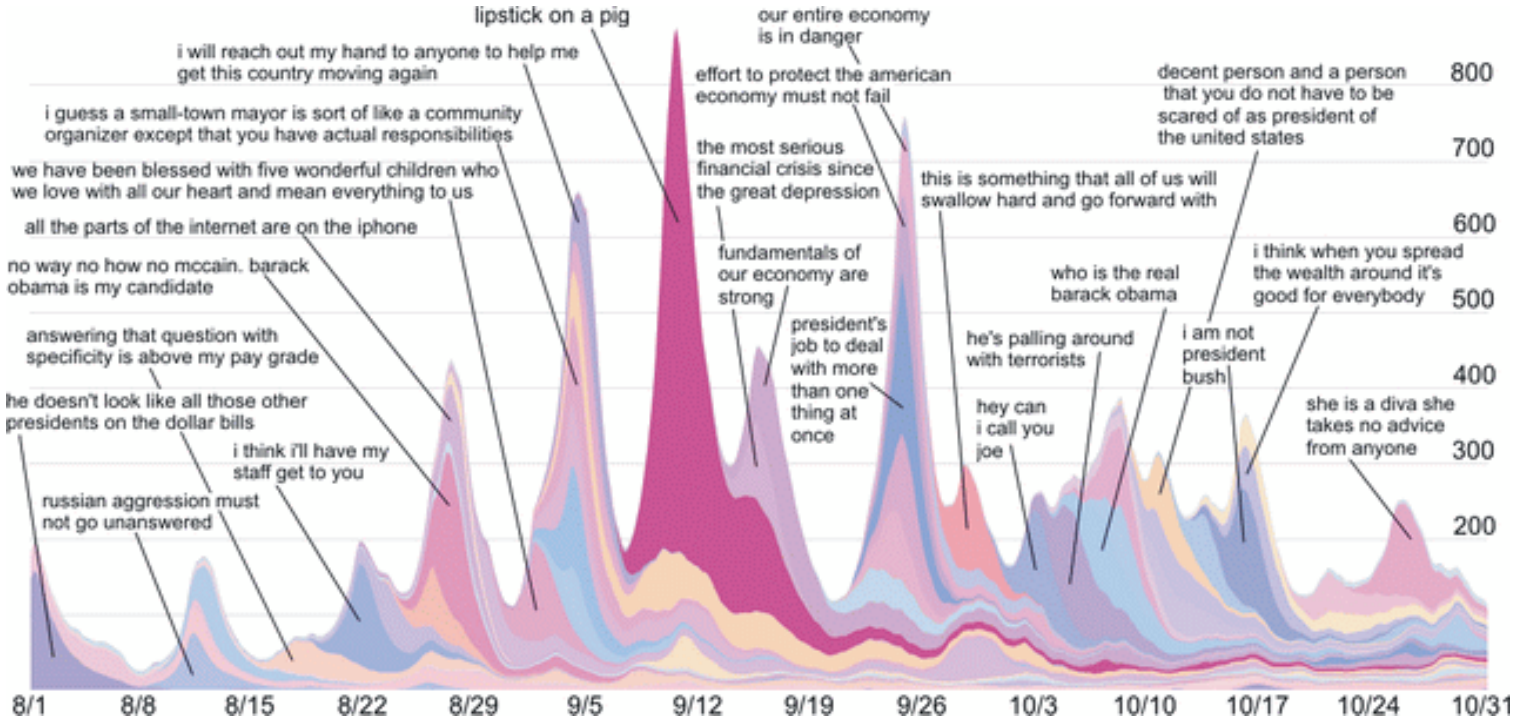
What society search for…

- Citizen response
- Feedbacks on political themes
- Campaigns
  - Why are people supporting a candidate?
- Law enforcement
  - Minority report

http://www.nytimes.com/2011/08/16/us/16police.html?_r=1

An example

http://memetracker.org



Most mentioned phrases in the 2008 US presidential campaign

How (Big) Social Network Data are actually produced

- Directly: By users that decide to share almost everything (photos, comments, political opinions, recipes, food recommendations, travel positions, mood…)
- Indirectly: By the social network providers that enrich the original user data, adding semantic meta data, statistics, usage patterns

Just to be sure they are worth to be named «real» Big Data…

- Volume: 3.48 Billion users and they post…
- Velocity: 5 Billion contents every (2-5 new users per second) day that have high…
- Variety: texts, images, videos, whose …
- Variability: is high and whose…
- Veracity: has to be checked, in order to get…
- Value: 20 Billion Dollar per year spent for social media advertising (Not too bad indeed…)

(Global Digital report)

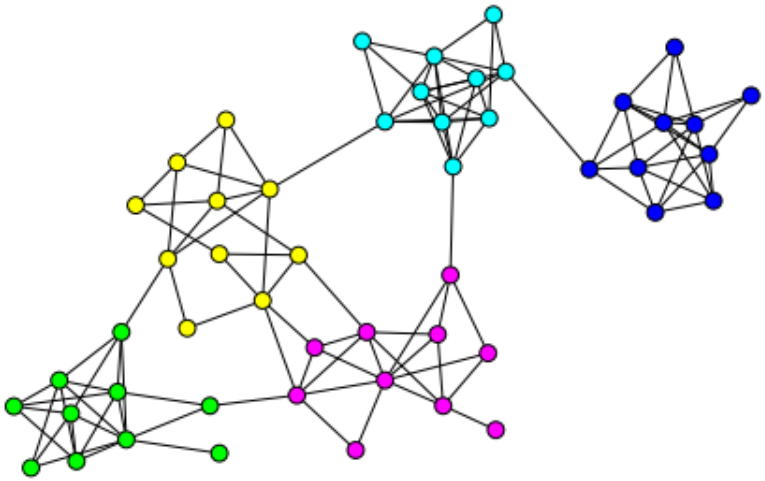
SN Bonus…

- Virality: reposting is an easy «cut and paste» for interesting information …
- Viscosity: they stick with users triggering reactions…
- Visualization: they intuitively make sense triggering (sometimes wrong ☺) decisions…

SN Structures

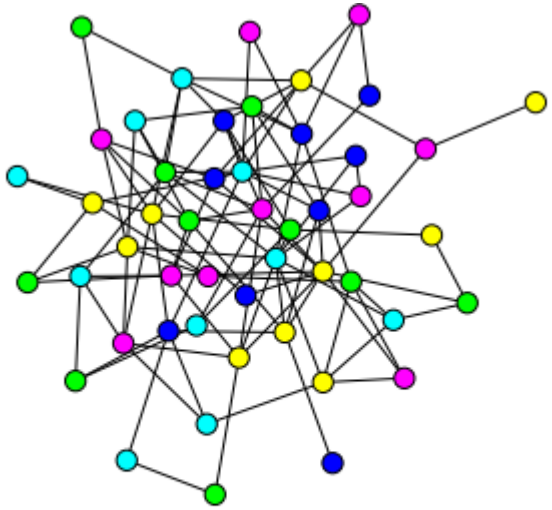
- They differ to other network structures (biological, transport and telecom to cite a few) because of the presence of positive degree correlations named as assortativity
  - *The perceived assortativity of social networks: Methodological problems and solutions, Fisher et al. (2017)*

In a word: similar tend to group together (homofily)



assortative
(edges within groups)

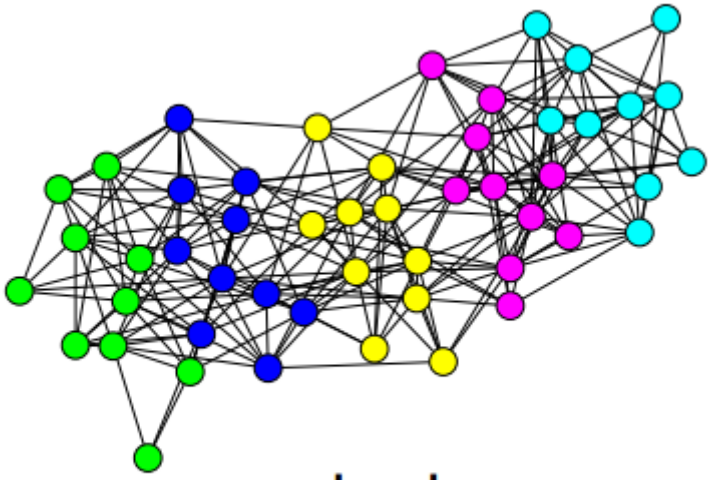
Few of them are diassortative



disassortative
(edges between groups)

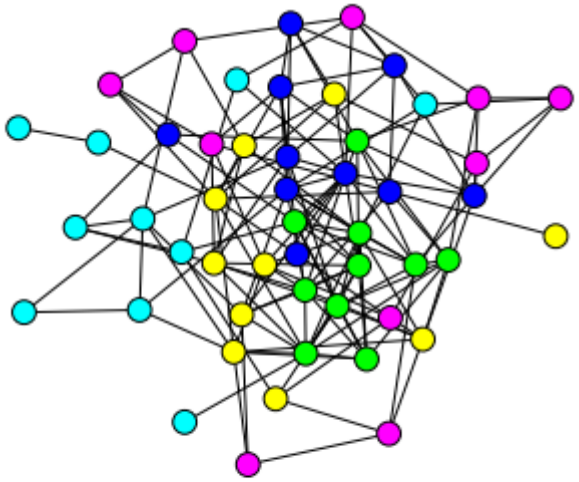
tinder

Some are hierarchical



ordered
(linear hierarchy of groups)

Few of them are core-periphery structured



core-periphery
(dense core, sparse periphery)

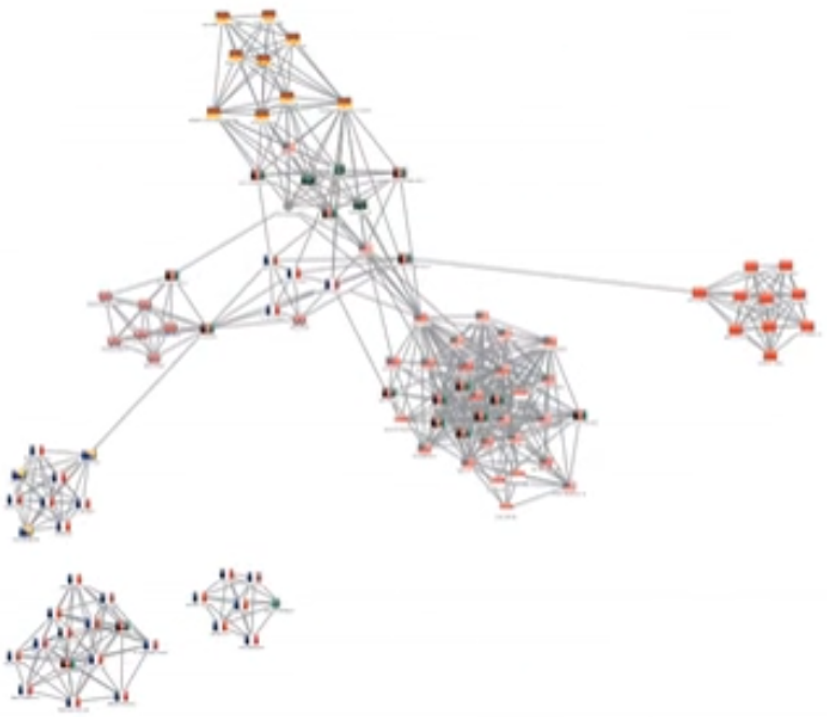
Zachary Karate Club

Networks of Network Scientists

Evaluating users relevance by their social ties

- Social ties evaluation dates back earlier than internet and social networks came on scene
- Two of the most cited works:
  - The strength of weak ties (Granovetter 1973)
  - Structural Holes: The Social Structure of Competition (Burt 1992)
- These papers outline the importance of centrality measures (specially betweenness centrality)
- Nowadays the actual size of networks pose many computational issues

SN metrics: Degree Centrality



Definition: Degree centrality assigns an importance score based purely on the number of links held by each node.

What it tells us: How many direct, 'one hop' connections each node has to other nodes within the network.

When to use it: For finding very connected individuals, popular individuals, individuals who are likely to hold most information or individuals who can quickly connect with the wider network.

A bit more detail: Degree centrality is the simplest measure of node connectivity. Sometimes it's useful to look at in-degree (number of inbound links) and out-degree (number of outbound links) as distinct measures, for example when looking at transactional data or account activity.

(Source: Cambridge Intelligence)

SN metrics: Betweenness centrality

Definition: Betweenness centrality measures the number of times a node lies on the shortest path between other nodes.

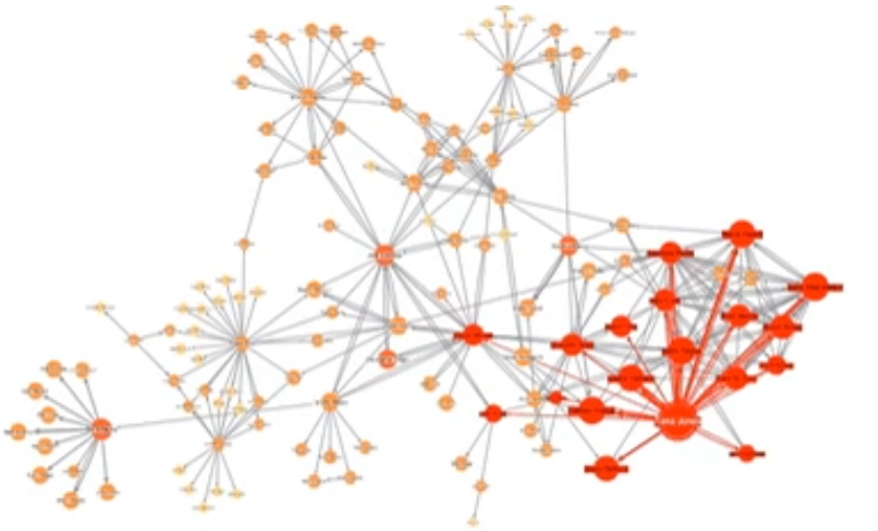
What it tells us: This measure shows which nodes act as 'bridges' between nodes in a network. It does this by identifying all the shortest paths and then counting how many times each node falls on one.

When to use it: For finding the individuals who influence the flow around a system.

A bit more detail: Betweenness is useful for analyzing communication dynamics, but should be used with care. A high betweenness count could indicate someone holds authority over, or controls collaboration between, disparate clusters in a network; or indicate they are on the periphery of both clusters.

(Source: Cambridge Intelligence)

SN metrics: Closeness centrality



Definition: This measure scores each node based on their 'closeness' to all other nodes within the network.

What it tells us: This measure calculates the shortest paths between all nodes, then assigns each node a score based on its sum of shortest paths.

When to use it: For finding the individuals who are best placed to influence the entire network most quickly.

A bit more detail: Closeness centrality can help find good 'broadcasters', but in a highly connected network you will often find all nodes have a similar score. What may be more useful is using Closeness to find influencers within a single cluster.

(Source: Cambridge Intelligence)

SN metrics: EigenCentrality

Definition: Like degree centrality, EigenCentrality measures a node's influence based on the number of links it has to other nodes within the network. EigenCentrality then goes a step further by also taking into account how well connected a node is, and how many links their connections have, and so on through the network.

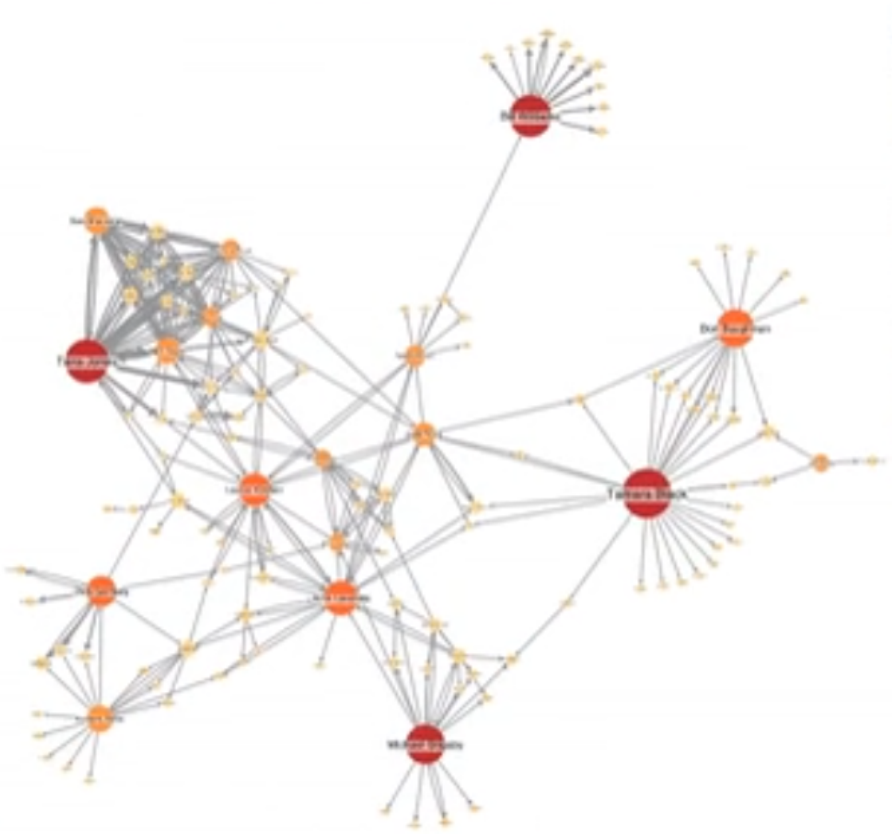
What it tells us: By calculating the extended connections of a node, EigenCentrality can identify nodes with influence over the whole network, not just those directly connected to it.

When to use it: EigenCentrality is a good 'all-round' SNA score, handy for understanding human social networks, but also for understanding networks like malware propagation.

A bit more detail: it is possible to calculates each node's EigenCentrality by converging on an eigenvector using the power iteration method.

(Source: Cambridge Intelligence)

SN metrics: PageRank



Definition: PageRank is a variant of EigenCentrality, also assigning nodes a score based on their connections, and their connections' connections. The difference is that PageRank also takes link direction and weight into account – so links can only pass influence in one direction, and pass different amounts of influence.

What it tells us: This measure uncovers nodes whose influence extends beyond their direct connections into the wider network.

When to use it: Because it factors in directionality and connection weight, PageRank can be helpful for understanding citations and authority.

A bit more detail: PageRank is famously one of the ranking algorithms behind the original Google search engine (the 'Page' part of its name curiously is the same of creator and Google founder, Larry Page).

(Source: Cambridge Intelligence)

Scalable computation of centrality measure

- For large evolving scalable graphs online computation of betweenness centrality has to be computed by network vertices and edges taking into account edge addition and removal
- In a recent paper a carefully engineered algorithm with out-of-core techniques tailored for modern parallel stream processing engines that run on clusters of shared-nothing commodity hardware showed satisfactory performances

Presented in: Scalable Online Betweenness Centrality in Evolving Graphs (Bonchi et al. 2016)

Scalable computation of centrality measure

Algorithm 1: Distributed computation for degree, closeness and betweenness centralities

- Initialization: for every $i \in \mathcal{V}$, compute $\mathcal{R}_i^1$ and $\mathcal{L}_i^1$ by local interactions;
- For $t < d_{\max}$. Given $\mathcal{R}_j^t$ and $\mathcal{L}_j^t$, which are obtained from local interactions with $j \in \mathcal{R}_i^1$ and $j \in \mathcal{L}_i^1$, respectively. Compute

$$\begin{aligned} \mathcal{R}_i^{t+1} &= (\cup_{j\in\mathcal{R}_i^1}\mathcal{R}_j^t) - \cup_{k=1}^t \mathcal{R}_i^k; \\ \mathcal{L}_i^{t+1} &= (\cup_{j\in\mathcal{L}_i^1}\mathcal{L}_j^t) - \cup_{k=1}^t \mathcal{L}_i^k; \end{aligned}$$

- Compute the cardinality of $\mathcal{R}_{i\to j}$ and $\mathcal{L}_{j\to i}$ by

$$|\mathcal{R}_{i\to j}| = \sum_{t=1}^{d_{\max}} |\mathcal{R}_j^t| \text{ and } |\mathcal{L}_{j\to i}| = \sum_{t=1}^{d_{\max}} |\mathcal{L}_j^t|.$$

Then, output the degree, closeness and betweenness centralities by

$$\begin{aligned} D_i &= |\mathcal{R}_i^1|; \\ C_i &= 1/\sum_{t=1}^{d_{\max}} (|\mathcal{R}_i^t| \cdot t); \\ B_i &= \sum_{j\in\mathcal{R}_i^1, k\in\mathcal{L}_i^1, j\neq k} |\mathcal{R}_{i\to j}| \cdot |\mathcal{L}_{k\to i}|. \end{aligned}$$

Presented in: Distributed Algorithms for Computation of Centrality Measures in Complex Networks (You et al 2016)

Social Network Analysis vs Social Media Analytics

- They are often (wrongly) considered the same…
- Different Goals
- Different Approaches
- Possible Sinergies

Social Media Analytics

- Target: Social Media Conversation
- Goal: Understanding user perception, sentiment and engagement w.r.t. a product
- Tasks: Brand Advocacy, Reputation Management, Community Management, Demand generation…

Social Network Analysis

- Target: User links
- Goal: Understanding user connectivity, centrality and relevance in a SN
- Tasks: Predictive Analysis for Link formation, Betwennes Centrality evaluation, Visual representation…

Outline

- A brief history of Social Networks (SN)
- The Big Data Challenges
- Social Networks (SN) Big Data Features
- **What happened so far**
- Conclusions

Dealing with Volume & Velocity

Let's take a look to the size of the data collection problem…

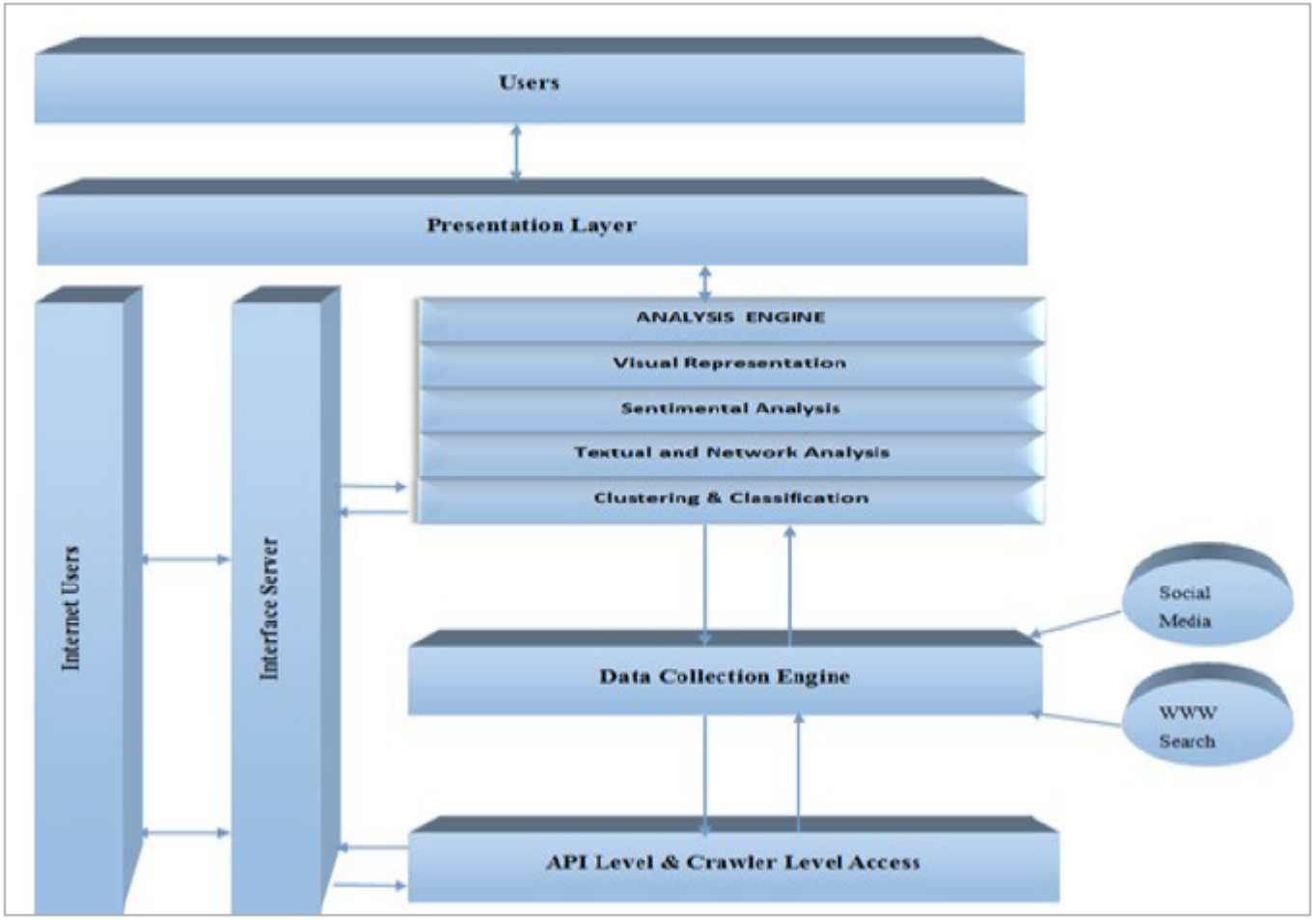
S. No.	Source of Data	No. of Elements	Frequency
1.	Tweets	More than 9000	Per Second
2.	Facebook Updates	More than 41000	Per Second
3.	Emails	2398534 Mails	Per Second
4.	Google Searches	More than 40000	Per Second
5.	YouTube	101604 Videos	Per Second
6.	Instagram	2000+ Photos	Per Second
7.	Tumblr	More than 1964 Posts	Per Second
8.	Skype	More than 1700 Calls	Per Second

(Source: Global Web Index)

Dealing with Volume & Velocity

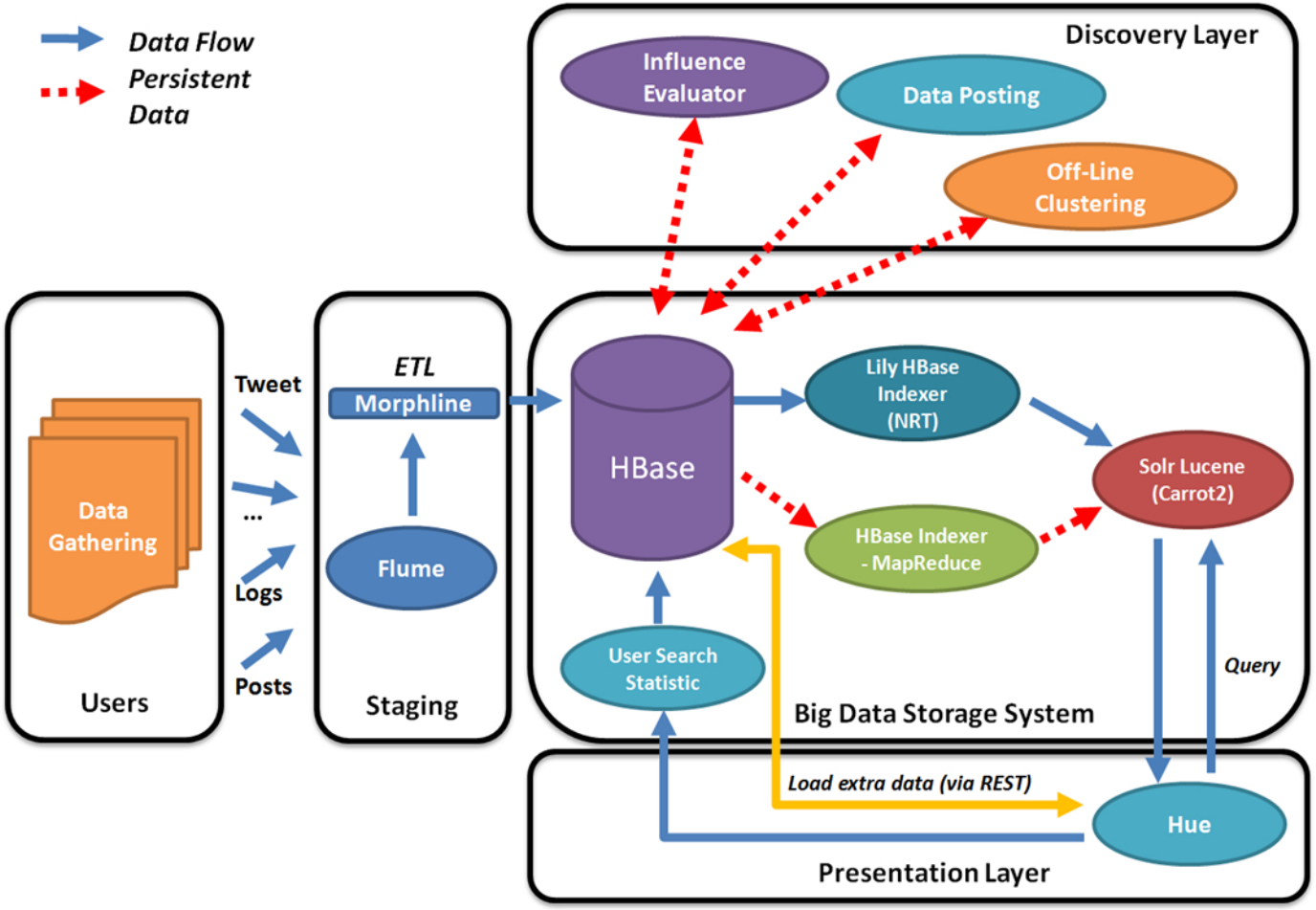
- Obviously there is a limit to the accessible data
- It's more a methodological problem rather than a «pure» research one (less fun unfortunately…)
- Basic tools we can use:
  - Interface Server
  - API level Access
  - Data Collection Engine
  - Analysis Engine

Dealing with Volume & velocity



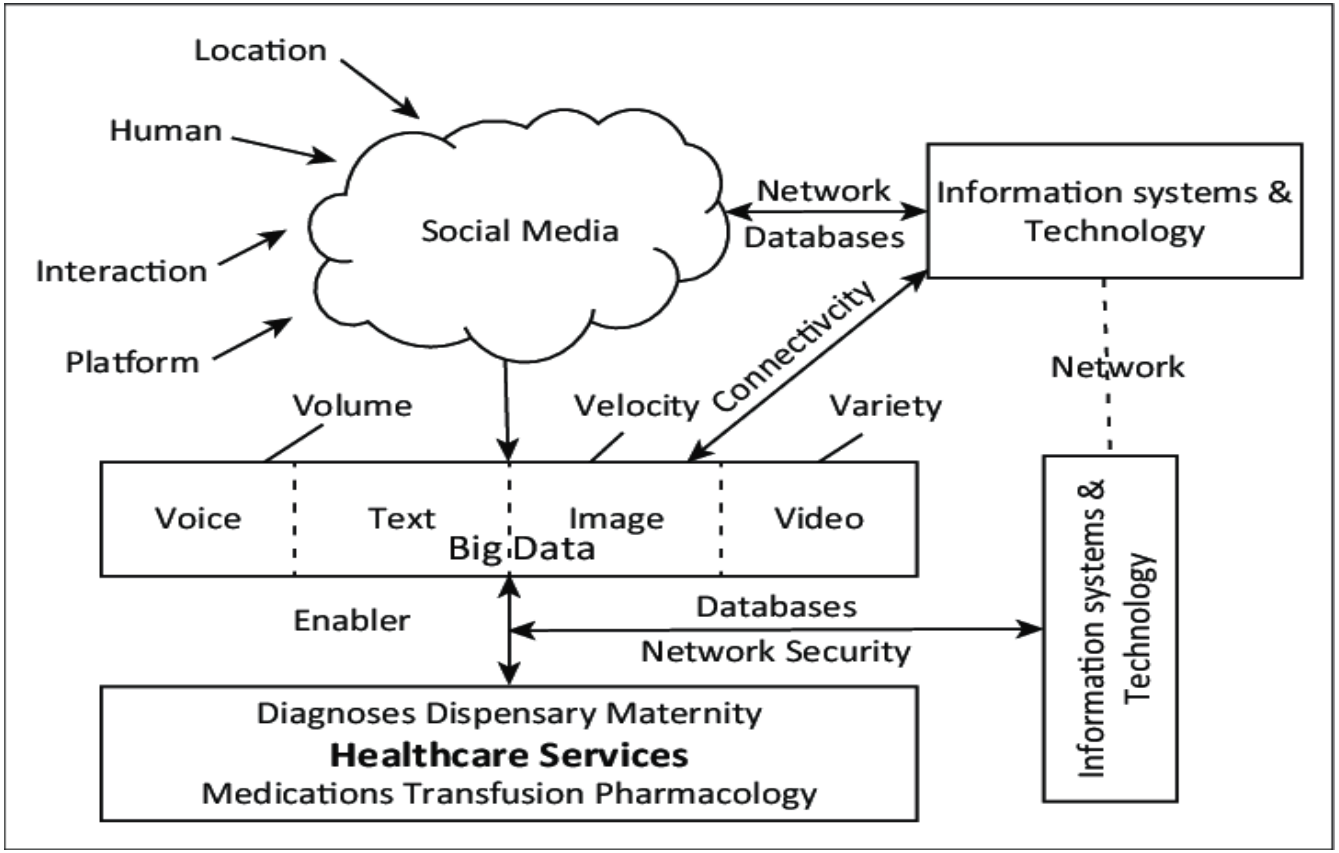
Proposed in: Data Collection And Analytics Strategies of Social Networking Websites (Kumar et al, 2015)

Dealing with Volume & Velocity



Proposed in: Discovering User Behavioral Features to Enhance Information Search on Big Data (Cassavia et al, 2017)

Dealing with Volume & Velocity



Proposed in: Information technology integration of social media with healthcare big data (Iyamu et al, 2018)

Social networks: The value of variety (Erickson 2003)

- People are healthier and happier when they have intimates who care about and for them
- But they also do better when they know many different people casually
- Acquaintance is important for better job (finding & performing) better than close ties
  - Remember “The strength of weak ties”…
- Wealthier people have diverse networks
- Acquaintance diversity also contributes to being better informed about health. People with wider networks are better informed about most things, but they may not realize how many of their good health practices go back to a thousand tiny nudges from casual conversations.

Visualizing variety

- In a sense variety can be considered as the set of “choice” individual can have
- To tackle heterogeneous features ontology can be leveraged
- Visualization is powerful
- Structural abstraction is used in Ontovis system that uses importance filtering to make large networks manageable and to facilitate analytic reasoning

Visualization of Movies (in orange) and People (in blue) Related to Woody Allen. The actors who worked most often with Woody Allen are Mia Farrow, Louise Lasser and Diane Keaton (Therefore, we conclude that Woody Allen often worked with his girlfriends on his movies…)

Presented in: Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction (Eliassi-Rad et al, 2006)

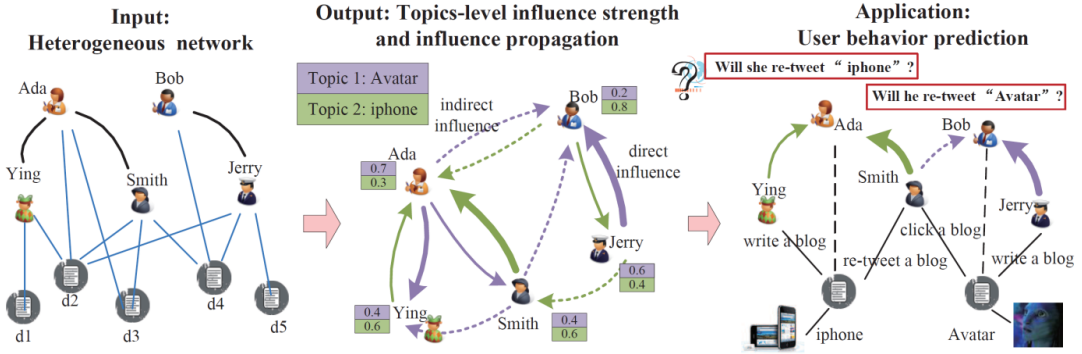
Variety in humans and cultural objects

- Correlation could emerge across the different activities a user can take part in
- A nice study on aNobii, a social platform with a world-wide user base of book readers, who like to post their readings, give ratings, review books and discuss them with friends and fellow readers has been analyzed
- Variety of roles: i) part social network, with user-to-user interactions, ii) part interest network, with the management of book collections, and iii) part folksonomy, with books that are tagged by the users
- Outcomes: user profiling cannot be reduced to considering just any one type of user activity (although important) but it is crucial to incorporate multiple dimensions to effectively describe users preferences and behavior
- Method: Experimental analysis carried out by means of Information Theory tools like entropy and mutual information suggests that tag-based and group-based profiles are in general more informative than wishlist-based ones

Proposed in: Analysis of a heterogeneous social network of humans and cultural objects (Agreste et al, 2014)

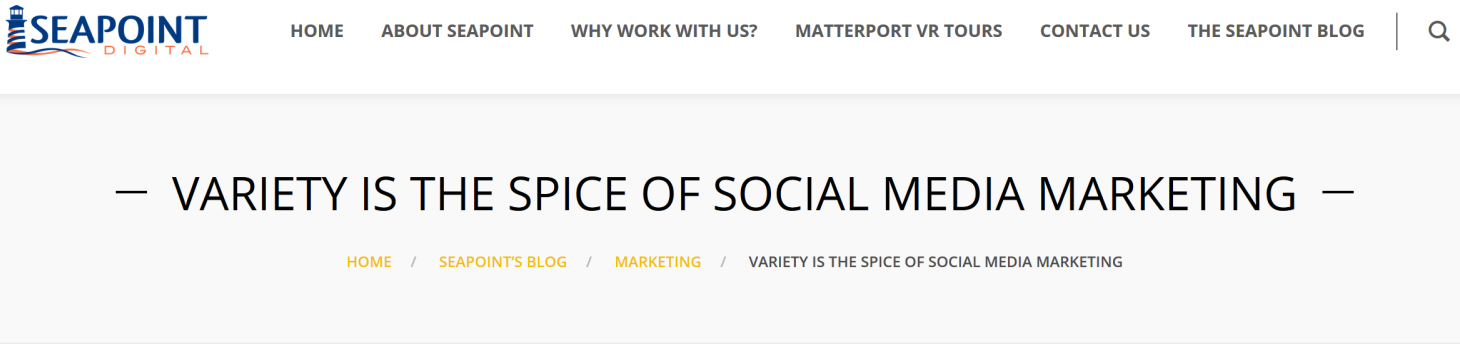
The effect of variety on social influence

- When social networks are heterogeneous (consisting of heterogeneous objects such as users, groups, and blogs), how the influence is affected by different types of objects on different topics (e.g., entertainment, marketing, and research)?
- Topic-level influence mining addressed by a generative model which utilizes both content and link information to mine direct influence strength in heterogeneous networks.
- Diffusion models for conservative and non-conservative influence propagations to learn indirect influence in social networks can be leveraged
- A study validated this approach in four different types of data sets: Twitter, Digg, Renren and Cora



Proposed in: Learning Influence from Heterogeneous Social Networks (Han et al, 2012)

Variety on Social Media



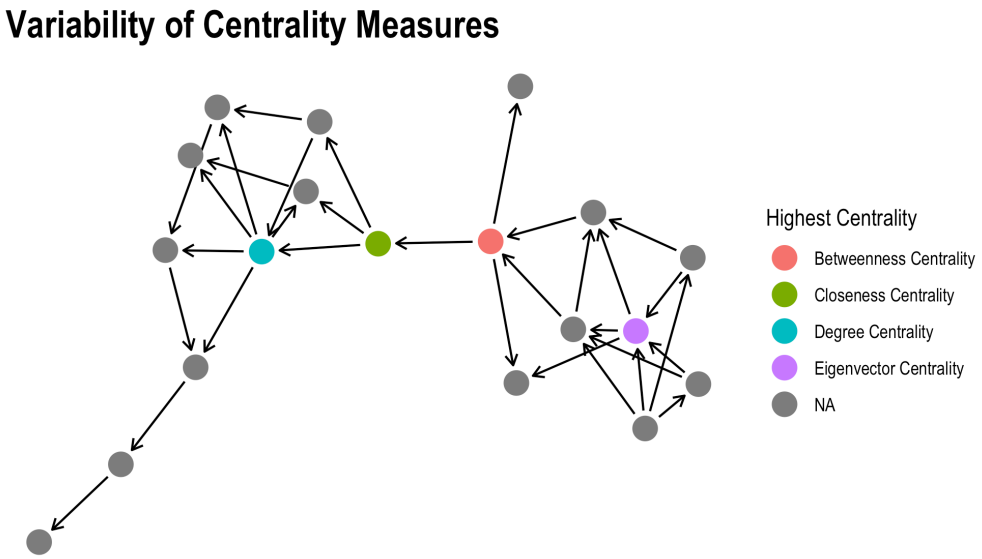
(https://seapoint.digital/variety-spice-social-media-marketing/)

Social networks: the role of variability

- Social groups present high variability
  - Age distribution
  - Node Fitness
  - Gender
  - Interests (they may change over time)
  - Personality
  - Geography
- The problem has to be tackled with a mix of sociology, economy, psicology skills aside the mere computer science ones
- Here scale is a limit to complete analysis

Social networks: the role of variability

- Centrality measures are themselves highly variable



(Source: Social Network Analysis with R)

Sources of variation in social networks (Atalay 2013)

- Why do some participants of social networks have so many contacts, while most others have so few?
- How important are age and randomness in explaining the variation in the number of contacts (i.e., the degree) that participants have?
- What is the underlying process that produces the degree distributions that are repeatedly observed in studies of social networks?
- Based on Jackson and Rogers framework, a model is built by allowing nodes to differ in the rate at which they can expect to gain additional links. The *fitness* of a node is defined as the probability that each of its meetings will generate a link based not only on in-cohort features.
- With more variability in fitness, there is more variability in the degree distribution of nodes of a particular age.

Analyzing Gender

- Scientists have established that social networks influence adolescents' substance use behavior, an influence that varies by gender. However, the role of gender in this mechanism of influence remains poorly understood. Particularly, the role an adolescent's gender, alongside the gender composition of his/her network, plays in facilitating or constraining alcohol use is still unclear.

- A study examined the associations among the gender composition of adolescents' networks, select network characteristics, intra-personal, inter-personal factors and alcohol use among a sample of adolescents in the US.

- They performed cross-sectional data from a 2010 study of 1,523 high school students from a school district in Los Angeles. Analyses of adolescents' network characteristics were conducted using UCINET 6; logistic regression analyses testing the associations between gender composition of the network and alcohol use were conducted using SPSS 20.

- The results indicate that the gender composition of adolescents' networks is associated with alcohol use. Adolescents in predominantly female or predominantly male friendship networks were less likely to report alcohol use compared to adolescents in an equal/balanced network. Additionally, depending upon the context/type of network, intrapersonal and interpersonal factors varied in their association with alcohol use.

Presented in: Adolescent Social Networks and Alcohol Use: Variability by Gender and Type (Jacobs et al, 2017)

Analyzing Geography

- Does geographical variability have potential implications on the structure of social networks?
- A study demonstrate that geographical variability produces large and distinctive features in the "social fabric" that overlies it
  - Many aggregate network properties can be fairly well predicted from relatively simple spatial demographic variables
  - Spatial variability exert substantial influence on network structure at the settlement level
  - Spatial heterogeneity induce substantial within network heterogeneity however geography drives many aggregate network properties in a predictable way

Presented in: Geographical variability and network structure (Jacobs et al, 2017)

Do we behave always the same?

- How the contextual expression of personality differs across interpersonal relationships?
- Participants in a study completed a five-factor measure of personality and constructed a social network detailing their 30 most important relationships
- Contextual personality ratings demonstrated incremental validity beyond standard global self-report in predicting specific informants' perceptions
- Variability in these contextualized personality ratings was predicted by the position of the other individuals within the social network. Across the studies, participants reported being more extraverted and neurotic, and less conscientious, with more central members of their social networks. Dyadic social network-based assessments of personality provide incremental validity in understanding personality, revealing dynamic patterns of personality variability unobservable with standard assessment techniques.

Presented in: Variability in Personality Expression Across Contexts: A Social Network Approach (Clifton, 2013)

The veracity problem

SOCIAL SCIENCE

The science of fake news

Addressing fake news requires a multidisciplinary effort

By **David Lazer, Matthew Baum, Yochai Benkler, Adam Berinsky, Kelly Greenhill, Filippo Menczer, Miriam Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven Sloman, Cass Sunstein, Emily Thorson, Duncan Watts, Jonathan Zittrain**

The rise of fake news highlights the erosion of long-standing institutional bulwarks against misinformation in the internet age. Concern over the problem is global. However, little is scientific questions raised by the proliferation of its most recent, politically oriented incarnation. Beyond selected references in the text, suggested further reading can be found in the supplementary materials.

WHAT IS FAKE NEWS?

We define "fake news" to be fabricated information that mimics news media content in form but not in organizational process or intent. Fake news outlets, in turn, lack the news media's editorial norms and processes for ensuring the accuracy and credibility of pernicious in that it is parasitic on standard news outlets, simultaneously benefiting from and undermining their credibility.

Our definition of fake news makes no assumptions about the characteristics of sources or amplification strategies. Some—notably First Draft and Facebook favor "false news" because of the use of fake news as a political weapon (*1*). We have retained it because of its value as a scientific construct, and because its political salience draws attention to an important subject.

THE HISTORICAL SETTING

Journalistic norms of objectivity and balance arose as a backlash among journalists against the widespread use of propaganda in World War I (particularly their own role in propagating it) and the rise of corporate public relations in the 1920s. Local and national oligopolies created by the dominant 20th cen-

(Science 2018)

The veracity problem

WORLD VIEW *A personal take on events*

CLARISSA SIMAS

The biggest pandemic risk? Viral misinformation

A century after the world's worst flu epidemic, rapid spread of misinformation is undermining trust in vaccines crucial to public health, warns **Heidi Larson**.

A hundred years ago this month, the death rate from the 1918 influenza was at its peak. An estimated 500 million people were infected over the course of the pandemic; between 50 million

disciplined and his article retracted 12 months after publication rather than 12 years, we might not be remarking that this year marks the twentieth anniversary of its publication.

(Nature 2018)

The veracity problem

MIT INITIATIVE ON THE DIGITAL ECONOMY RESEARCH BRIEF

THE SPREAD OF TRUE AND FALSE NEWS ONLINE

By Soroush Vosoughi, Deb Roy, and Sinan Aral

FALSE NEWS IS BIG NEWS.

Barely a day goes by without a new development about the veracity of social media, foreign meddling in U.S. elections, or questionable science.

RESEARCH HIGHLIGHTS

We investigated the differential diffusion of all the verified, true and false news stories distributed

(MIT 2018)

The veracity problem

Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories. The effects were most pronounced for false political news than for news about terrorism, natural disasters, science, urban legends, or financial information.

Controlling for many factors, false news was 70% more likely to be retweeted than the truth.

Novelty is an important factor. False news was perceived as more novel than true news, which suggests that people are more likely to share novel information.

T 2018)

The veracity problem

- The large availability of user provided contents on online social media facilitates people aggregation around shared beliefs, interests, worldviews and narratives
- In spite of the enthusiastic rhetoric about the so called collective intelligence unsubstantiated rumors and conspiracy theories—e.g., chemtrails, reptilians or the Illuminati—are pervasive in online social networks
- A study examined on a sample of 1.2 million of individuals, how information related to very distinct narratives—i.e. main stream scientific and conspiracy news—are consumed and shape communities on Facebook
  - Conspiracy theories and scientific news generates homogeneous and polarized communities (i.e., echo chambers) having similar information consumption patterns
- The study measure how users respond to troll information—i.e. parodistic and sarcastic imitation of conspiracy theories:
  - 77.92% of likes and 80.86% of comments are from users usually interacting with conspiracy stories;

Presented in: The spreading of misinformation online (Bessi et al, 2016)

A critical point of view on research

SYMPOSIUM/7

SOCIAL MEDIA RESEARCH AFTER THE FAKE NEWS DEBACLE

Richard Rogers
University of Amsterdam

1. Introduction: The coming crisis in social media research

The purpose of the following is to reintroduce contemporary critiques of social media research, as they are gathering steam following the 'fake news debacle', which I come to. These are not social media or platform critiques per se, such as platformization which refers to how the web is becoming enclosed and overwritten by social media (Helmond 2015). Embedded in the research critique is some discussion of Facebook policy as well as Twitter rules (for example), but that is not the main effort here.

A critical point of view on research

Regarding the crisis, it has been argued that unlike other disciplines computer science has not had the 'reckoning' that chemistry had after dynamite and poison gas, physics after the nuclear bomb, human biology after eugenics, civil engineering after bridge, dam and building collapses, and so forth (Zunger 2018)

Fake news mitigation

- Fake news mitigation aims to reduce the negative effects brought by fake news
- From a network analysis perspective, the goal is to minimize the scope of fake news spreading on social media
- Steps:
  - Key spreaders of fake news need to be discovered such as provenances and persuaders;
  - The potential population affected by a fake news has to be estimated for decision-makers to mitigate otherwise influential fake news;
  - Identify specific users to block the cascade of fake news, and even to start mitigation campaigns to immunize users are required to minimize the influence of fake news.

Presented in: Studying Fake News via Network Analysis: Detection and Mitigation (Berneard et al, 2019)

News Automation

- To deal with the sheer volume of information and gain competitive advantage, the news industry has started to explore and invest in news automation
- Reuters Tracer is a system that automates end-to-end news production using Twitter data
  - It detects, classify, annotate, and disseminate news in real time for Reuters journalists without manual intervention
- Tracer is topic and domain agnostic
- It leverages a bottom-up approach to news detection, and does not rely on a predefined set of sources or subjects:
  - It identifies emerging conversations from 12+ million tweets per day and selects those that are news-like
  - It contextualizes each story by adding a summary and a topic to it, estimating its newsworthiness, veracity, novelty, and scope, and geotags it

Presented in: Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data (Liu et al, 2017)

Truth Discovery

- Three important challenges for truth discovery:
  - Misinformation spread;
  - Data sparsity;
  - Scalability
- Big data social media sensing applications
  - Social sensing applications often generate large amounts of data during important events (e.g., disasters, sports, unrests)
- Scalable Robust Truth Discovery (SRTD) scheme explicitly considers various source behaviors, content analysis of claims, and historical contributions of sources in a holistic truth discovery solution
- It is a light-weight distributed framework to implement the SRTD scheme and improve computational efficiency

Presented in: On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications (Zhang et al, 2018)

Online Reputation Management

- This task according to several studies is:
  - Necessary;
  - Unpleasant;
  - Disempowering
- The main problem is that repairing a reputation damage is considered almost impossible:
  - **Ignoring it… is unsatisfactory**
  - **Persuasion… is ineffective**
  - **Rebuttal… is risky**
  - **Asking for help… fails because no one is responsible**

Presented in: Necessary, Unpleasant, and Disempowering: Reputation Management in the Internet Age (Woodruff 2014)

Some interesting examples

- Case 1: When one of the world's biggest air carriers, United Airlines, refused to compensate a passenger who was a professional musician for breaking his $3,500 guitar in 2008, he eventually wrote a song about his lengthy but failed negotiations with the company. Then he sang the song on a derogatory music video posted on YouTube in 2009. His protest video ''United Breaks Guitars'' was seen by millions of people in a matter of days, and as a result the case received widespread coverage in both Internet media – blogs, forums and news websites – as well as print and TV. Reacting to the groundswell of adverse publicity, the carrier quickly responded with a settlement offer.

Presented in: Social media, reputation risk and ambient publicity management (Aula, 2010)

Some interesting examples

- Case 2: Clothing company H&M became the subject of an unexpected scandal in New York after a student found bags of its unsold clothes that had been mutilated and dumped in the garbage by store personnel. Shocked that the store trashed the clothes instead of donating them to nearby agencies that would have distributed them to the needy, the student informed the New York Times. When questioned by reporters H&M store representatives were caught off guard and refused to comment. Soon the story found its way onto Twitter, the micro-blogging service. After public outrage quickly spread via social media the company gave their first statement about the ‘‘trashgate’’ incident.

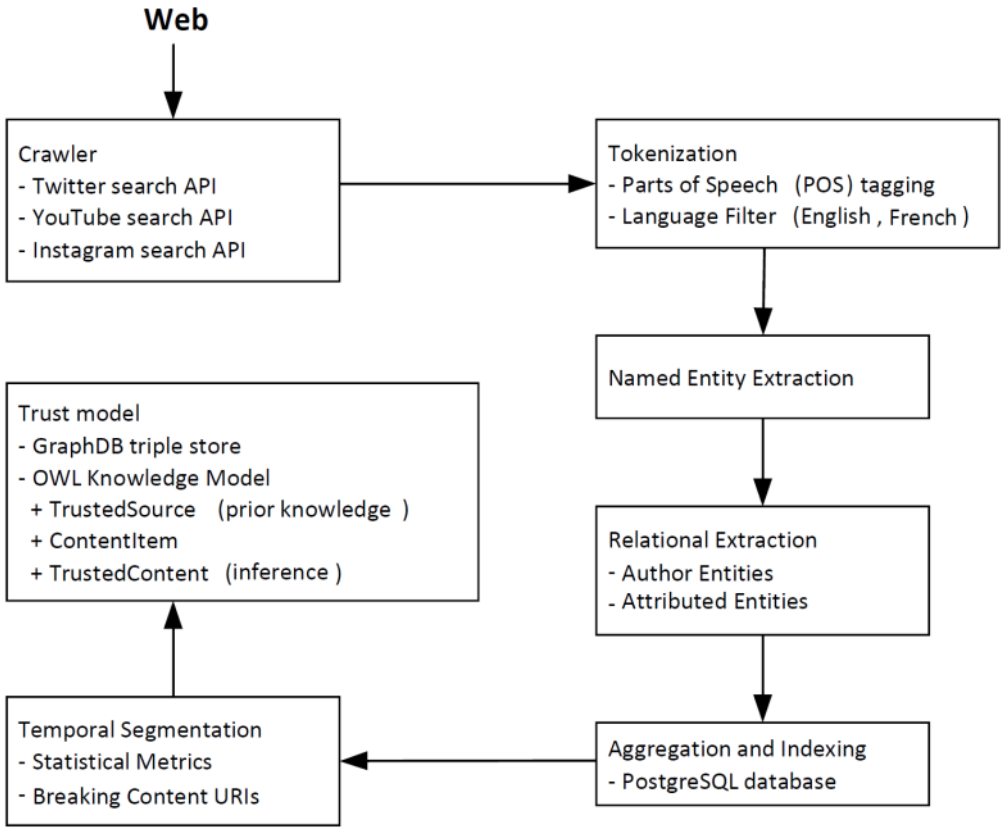
Presented in: Social media, reputation risk and ambient publicity management (Aula, 2010)

Some interesting examples

- Case 3: A car dealership in Finland found itself in an awkward situation when a customer read an extremely insulting description of himself among some internal documents. The incensed customer wrote about what had happened on an Internet chat forum, and from there the story spread to the tabloid newspapers. Not only the car dealership, but also officials of the importer of the car brand were pressed by the general media for comment after there were indications that the incident would affect sales.

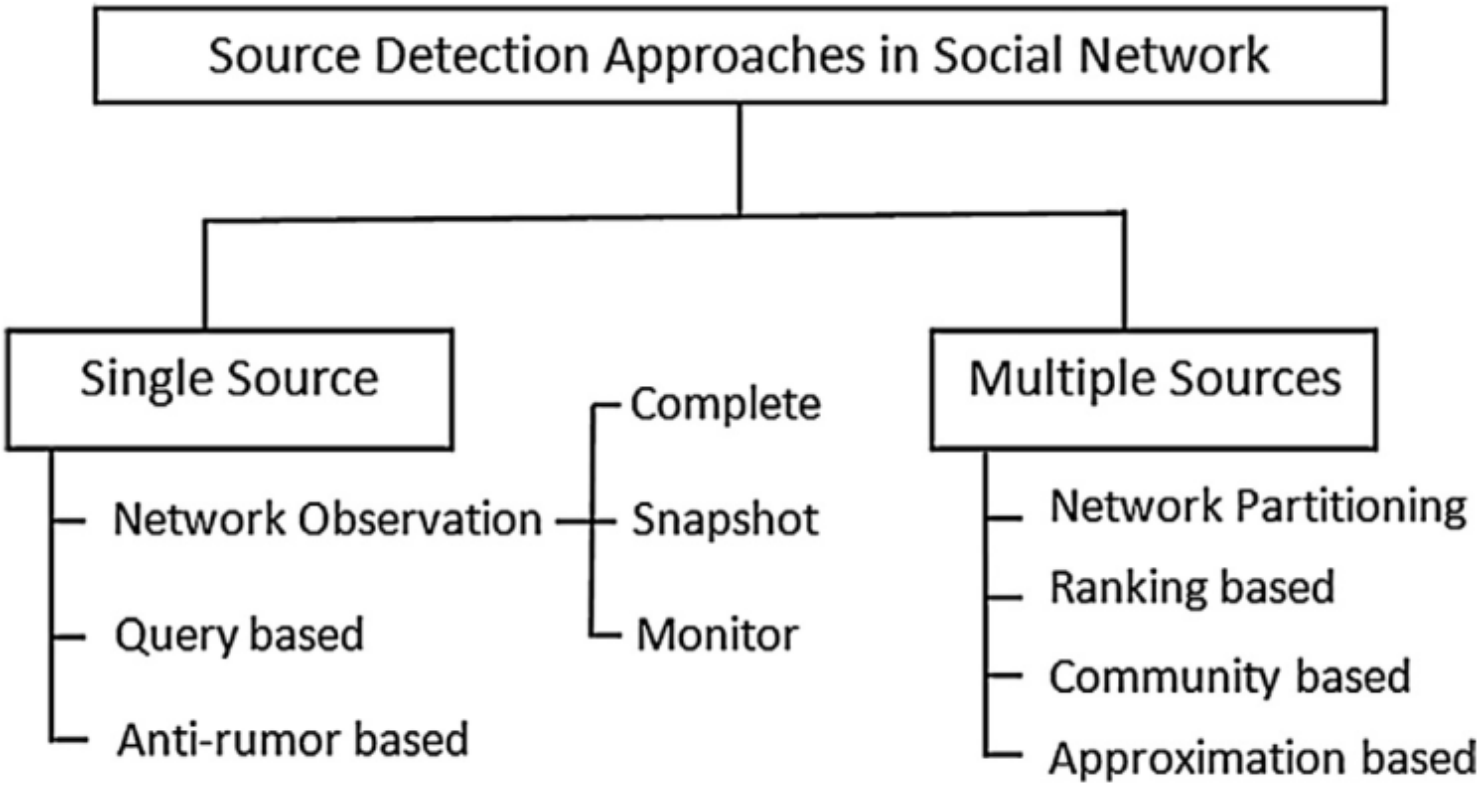
Presented in: Social media, reputation risk and ambient publicity management (Aula, 2010)

Breaking News Checking



Presented in: Veracity and Velocity of Social Media Content during Breaking News: Analysis of November 2015 Paris Shootings (Middleton et al, 2016)

Source detection



Presented in: Source detection of rumor in social network –A review (Attar et al, 2019)

How to extract value

- The large body of user generated content (UGC) is a continuous source of profitable big data that can be used for:
  - brand advocacy
  - reputation management
  - competitor analysis
  - community management
  - customer management
  - viral marketing
  - sentiment analysis
  - community detection
  - influence spread
  - user recommendation

How to extract value

- All these approaches can be used for gathering crucial information like:
  - Why people like or dislike a product?
  - Who are the top competitors online?
  - What are the media that mainly deliver interesting information about a company?
  - What are the features associated to a given brand?
  - What are the sentiment about a political candidate?
  - What are the trend topics?
  - Which users should be involved in product information spreading?
  - How trust evolves in a community?
  - How communities interact?

The effect of UGC on marketing

- UGC comes from a variety of venues, such as tweets or Facebook pages, pictures (Pinterest), blogs, microblogs, and product reviews (Amazon, Yelp)
- Empirical findings show that UGC has significant effects on brand images, purchase intentions, and sales (An important role is played by '@' and "#")
- Due to their unstructured nature it is important to leverage clustering and probabilistic topic modeling to properly identify sentiments that have a marketing value
- One of the most effective social in this respect is Twitter

The effect of UGC on marketing

- Useful questions to answer:
  - RQ1: What brand-related topics do consumers discuss on Twitter?
  - RQ2: What are the rankings of brand sentiments within and across industries?
  - RQ3: How can we identify specific product and service issues that consumers complain about?
  - RQ4: How can we identify the merits of products and services that consumers feel good about?
  - RQ3and RQ4 relates to company specific features and are not included in this tutorial

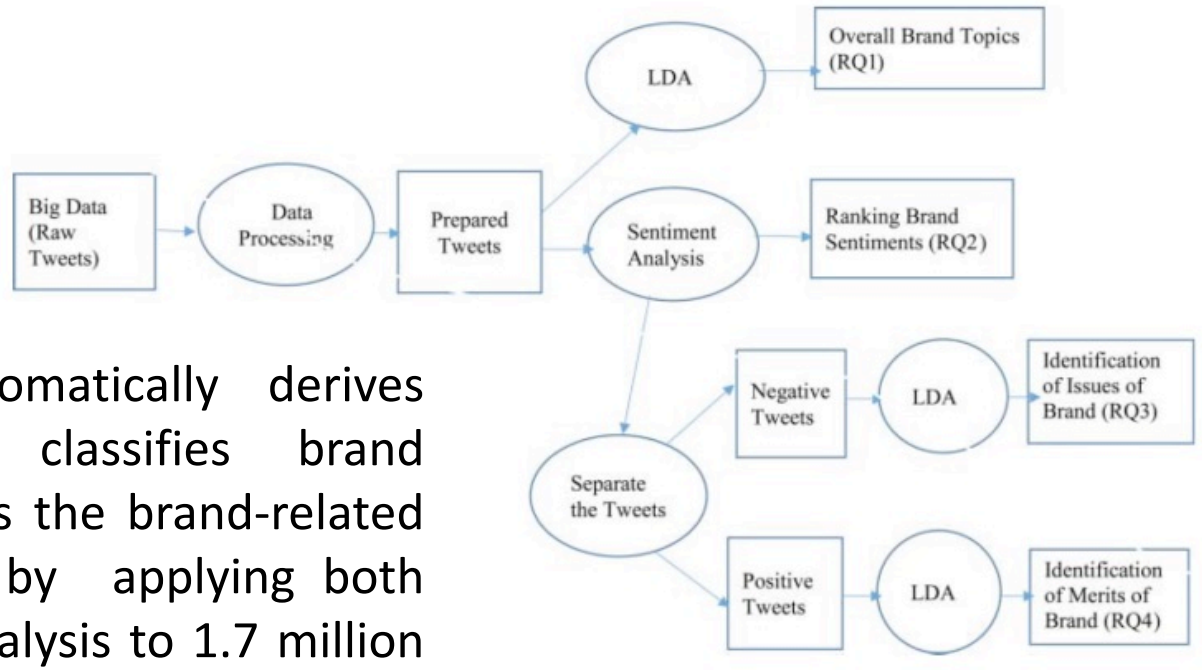
Presented in: An Investigation of Brand-Related User-Generated Content on Twitter (Liu et al, 2017)

The effect of UGC on marketing

Text Mining and Sentiment Analysis in Marketing Literature

Method	Validation
Lee and Bradlow (2011) analyze Epinions reviews for digital cameras and automatically extract attributes of products and relative positions of brands.	Buying guides, lab survey, and correspondence analysis
Netzer et al. (2012) use messages from Edmunds and a drug forum and extract market structure with text mining and network analysis from messages.	Survey and sales data
Moon, Park, and Seog Kim (2014) use information for 121 movies and 9,380 IMDb reviews and combine clusters of text reviews with rating scores to predict box-office sales.	Increases in the explanatory power of sales models for box office
Schweidel and Moe (2014) analyze 7,565 posts with sentiment analysis.	Linguistic inquiry, word count, and sampled and coded 200 reviews
Tirunillai and Tellis (2014) analyze 350,000 reviews with LDA and extract product dimensions and valence.	Human raters (face validity); *Consumer Reports* (external validity)
Homburg, Ehm, and Artz (2015) analyze 115,000 online messages with sentiment analysis.	Field experiment (internal validity)
Ma, Sun, and Kekre (2015) use a panel data of tweets from 714 customers and classify tweets as negative, neutral, or positive with text mining.	No validation for text mining results

The effect of UGC on marketing



This framework automatically derives brand topics and classifies brand sentiments. It explores the brand-related questions on Twitter by applying both LDA and sentiment analysis to 1.7 million tweets, moreover they leverage benchmarking against ACSI and expertise from industry experts

RQ1: What brand-related topics do consumers discuss on Twitter?

User-Generated Content Topics: Industry Similarities and Differences

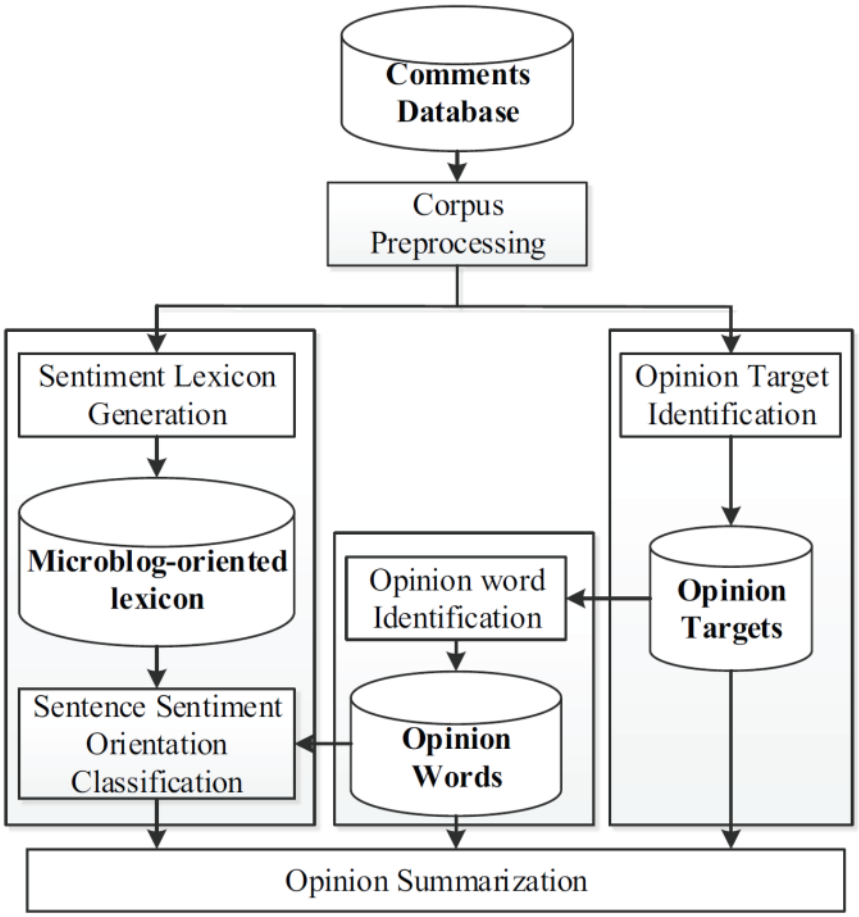
Topics/Industry	Fast Food	Department Store	Footwear	Telecommunications	Electronics	Average
Product	47.9%	23.5%	55.9%	45.6%	70.6%	48.7% (17.2%)
Service	20.4%	29.6%	1.8%	28.5%	8.0%	17.6% (12.4%)
Promotion	15.7%	24.6%	15.3%	5.0%	5.8%	13.3% (8.1%)
Competitors	4.8%	5.4%	7.7%	5.7%	9.0%	6.5% (1.8%)
News/trends	6.7%	9.2%	9.2%	2.6%	5.1%	6.6% (2.8%)

RQ2: What are the rankings of brand sentiments within and across industries?

Brand Sentiments Across Industries

Industry	Company	Negative	Neutral	Positive	Negative (%)	Positive (%)
Fast-food restaurant	Burger King	38,123	36,633	12,071	43.9	13.9
	KFC	27,105	20,093	7,422	49.6	13.6
	McDonald's	108,686	81,698	35,693	48.1	15.8
	Wendy's	26,493	23,975	10,190	43.7	16.8
	Industry average				**46.3**	**15.0**
Department store	JCPenney	10,589	8,558	3,937	45.9	17.1
	Kohl's	9,103	9,971	6,187	36.0	24.5
	Macy's	21,980	20,155	9,919	42.2	19.1
	Sears	8,537	6,007	2,972	48.7	17.0
	Industry average				**43.2**	**19.4**
Footwear	Adidas	21,870	16,450	7,853	47.4	17.0
	New Balance	7,218	8,859	4,968	34.3	23.6
	Nike	55,980	40,595	22,383	47.1	18.8
	PUMA	6,953	7,534	3,928	37.8	21.3
	Industry average				**41.7**	**20.2**
Telecommunications	Comcast	70,280	24,959	10,050	66.7	9.5
	Cox	10,111	4,531	2,096	60.4	12.5
	Dish	16,015	11,056	2,327	54.5	7.9
	TWC	27,337	11,549	4,757	62.6	10.9
	Industry average				**61.1**	**10.2**
Electronics	LGUS	6,534	4,027	2,663	49.4	20.1
	Panasonic	950	1,079	528	37.2	20.6
	Samsung	5,024	3,423	2,045	47.9	19.5
	Sony	55,329	31,469	18,774	52.4	17.8
	Industry average				**46.7**	**19.5**

OPINION MINING



Presented in: Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-aware Approach (Lin et al, 2014)

Social Network Formation

- Social network formation have attracted increasing attention from both physical and social scientists
- Network embedding algorithms in machine learning literature consider broad heterogeneity among agents while the social sciences emphasize the interpretability of link formation mechanisms
- A social network formation model that integrates methods in multiple disciplines and retain both heterogeneity and interpretability can be built by leveraging "endowment vectors" that encapsulates agents features and game-theoretical methods to model the utility of link formation

Presented in: An interpretable approach for social network formation among heterogeneous agents (Yuan et al, 2018)

Business Processes and SN

- APQC's Process Classification Framework (PCF) serves as a high-level, industry-neutral enterprise process model that allows organizations to see their business processes from a cross-industry viewpoint

Presented in: Social Network Analysis and Mining for Business Applications (Bonchi et al, 2011)

Business Processes and SN

Table I. Operating Processes

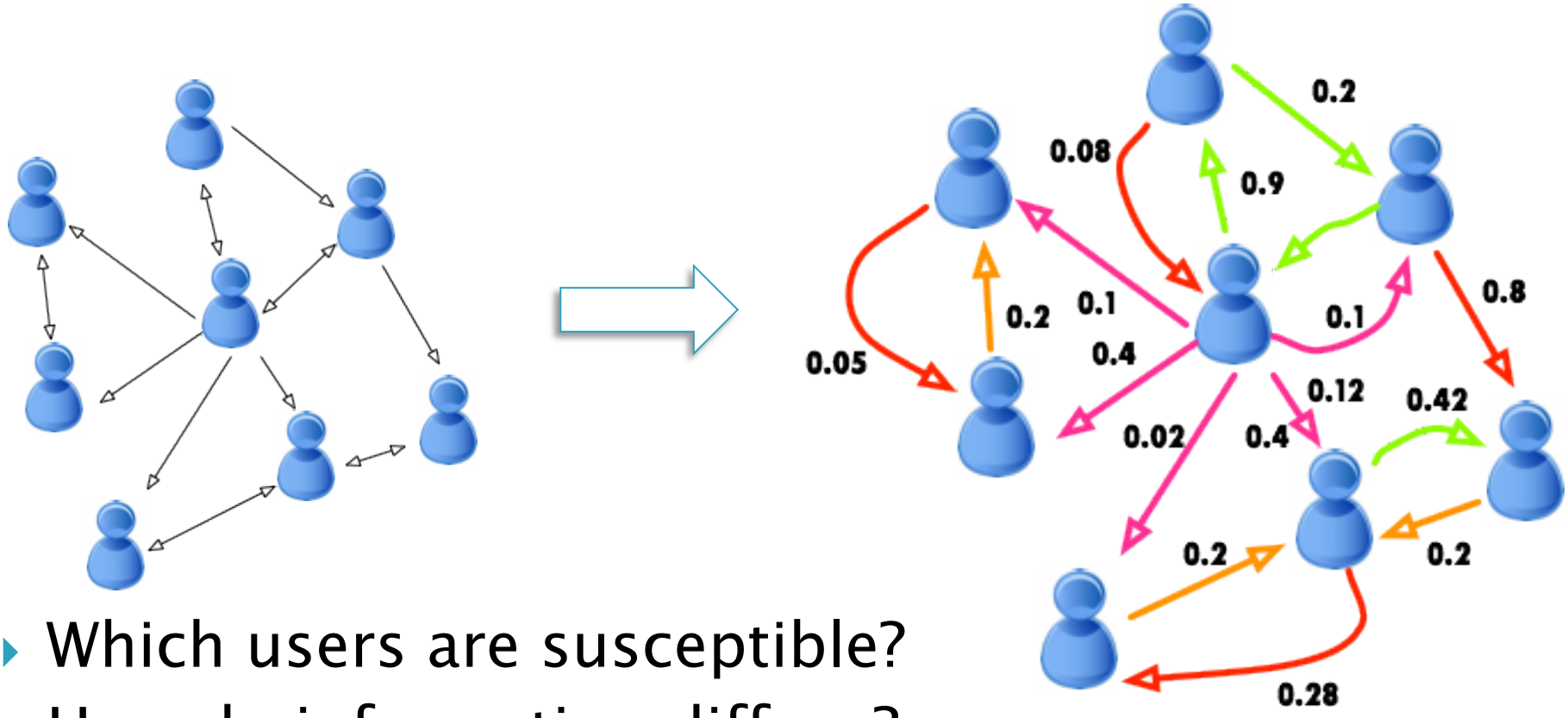
Process Category	*Process group or Activity*	*Technical area*

Table II. Management and Support Processes

Process Category	*Process group or Activity*	*Technical area*
6. Human Capital	Internal social networking Professional development Recruiting	SN Support tools Expert routing Social search
7. Information Technology	Resource allocation Information sources Content Management	Measurement Data preparation Privacy
8. Financial Resources	Customer & product strategies Customer-product mix Manage internal controls	SN Mining Community Community
9. Property Management	N.A.	N.A.
10. Environmental issues	N.A.	N.A.
11. External Relationships	Public relations program Legal and ethical issues Social networking	Monitoring Privacy SN support tools
12. Knowledge Management	Knowledge sharing Strategic KM	Internal social networks SN Mining

Presented in: Social Network Analysis and Mining for Business Applications (Bonchi et al, 2011)

Influence and Authority

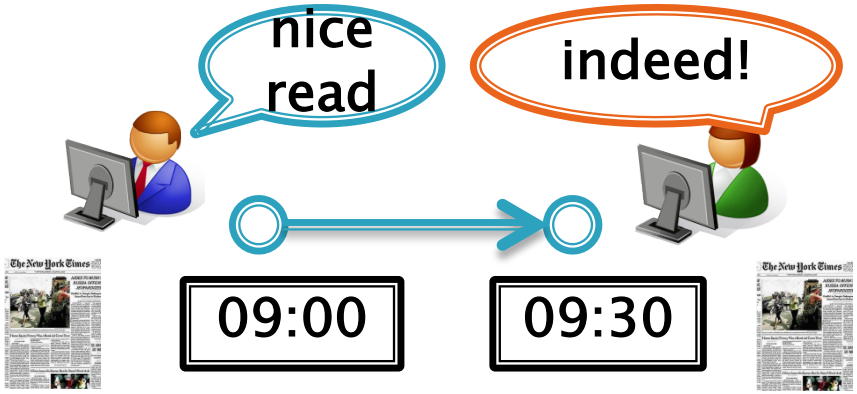


- Which users are susceptible?
- How do information diffuse?
- Can we trust a user/an opinion?

Smart models

- Logistic/Linear regression
- Deep Networks
- SVD and other MF models
- Restricted Bolzman machines
- Markov Chains
- Clustering
- Topic Models
- Boosted Decision Trees
- Associations

Influence in on-line social networks



users perform actions

post messages, pictures, video

buy, comment, link, rate, share, like, retweet

users are connected with other users

interact, influence each other

actions propagate

Credits: Giuseppe Manco

Influence or Homophily?

Homophily

tendency to stay together with people similar to you

"Birds of a feather flock together"

Social influence

a force that person A (i.e., the influencer) exerts on person B to introduce a change of the behavior and/or opinion of B

Influence is a causal process

Problem: How to distinguish social influence from homophily and other factors of correlation

Some relevant sources:

"Feedback Effects between Similarity and Social Influence in Online Communities" (Kleinberg et al, 2008)

"Influence and correlation in social networks" (Kumar et al, 2008)

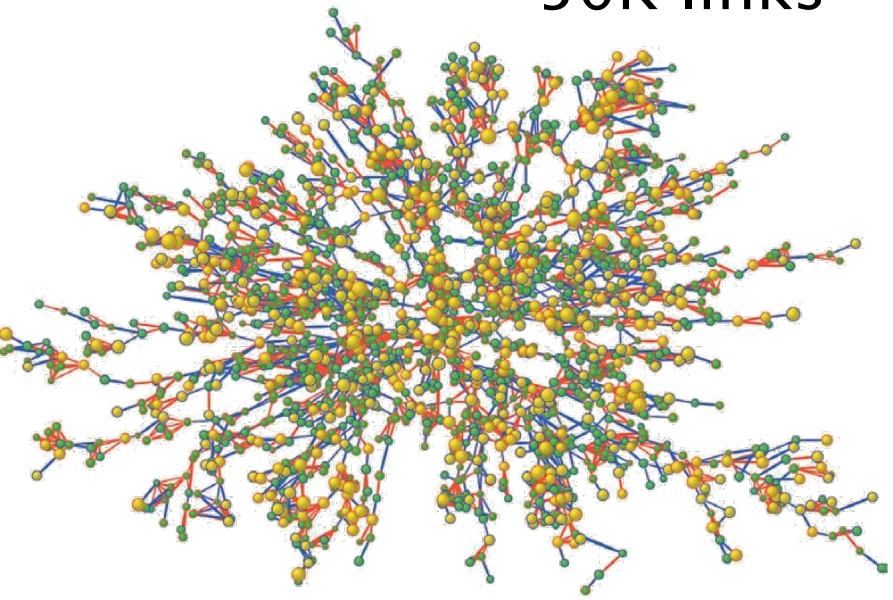
"Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks" (Aral et al, 2009)

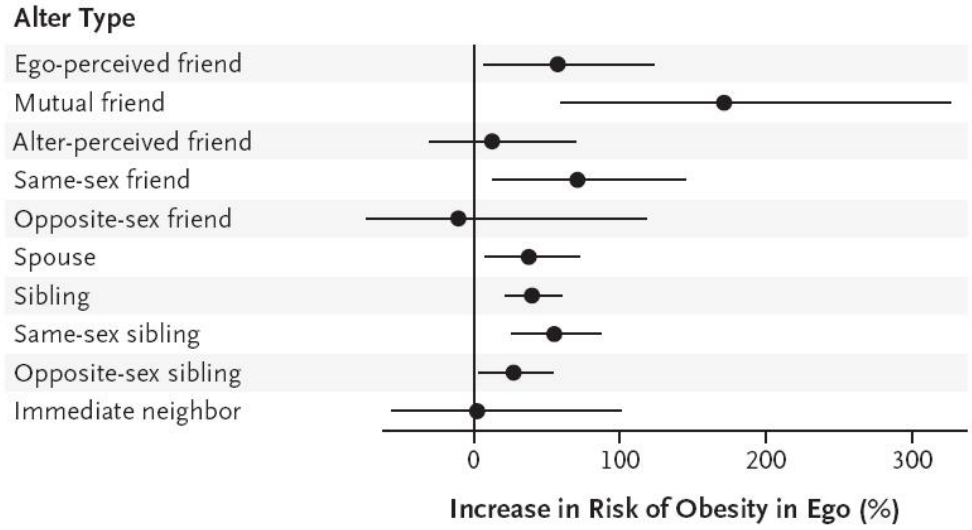
"Information Diffusion and External Influence in Networks" (Leskovec et al, 2012)

The Spread of Obesity in a Large Social Network over 32 Years

Christakis and Fowler, New England Journal of Medicine, 2007

Data set: 12,067 people from 1971 to 2003, 50K links



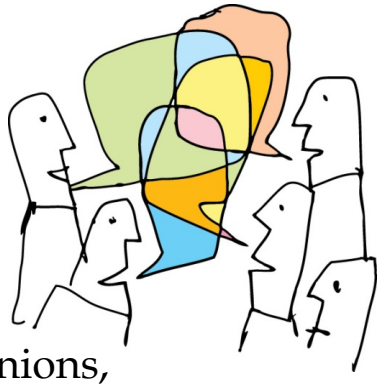


Obese Friend → 57% increase in chances of obesity

Obese Sibling → 40% increase in chances of obesity

Obese Spouse → 37% increase in chances of obesity

Social Influence and Viral Marketing



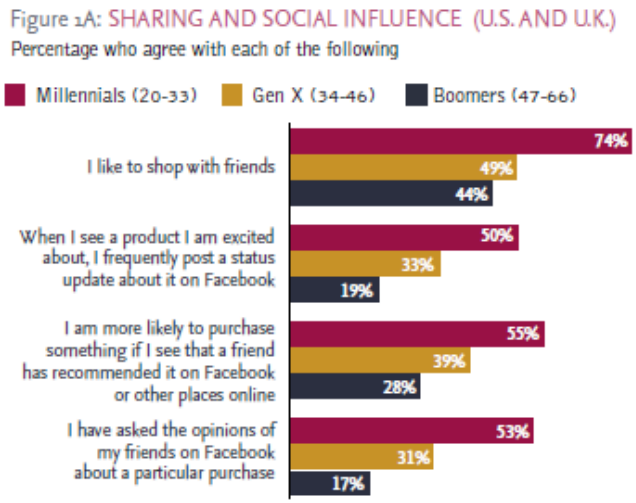
IDEA: exploit social influence for marketing

Basic assumption: word-of-mouth effect, thanks to which actions, opinions, buying behaviors, innovations and so on, propagate in a social network.

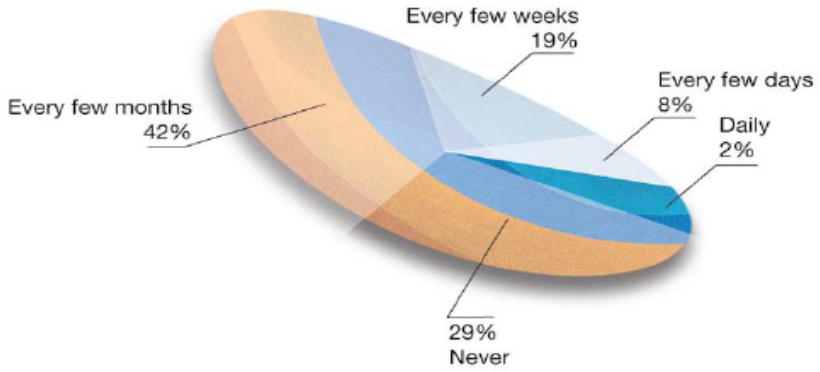
Target users who are likely to produce word-of-mouth diffusion, thus leading to additional reach, clicks, conversions, or brand awareness

Target the influencers

Sharing and social influence



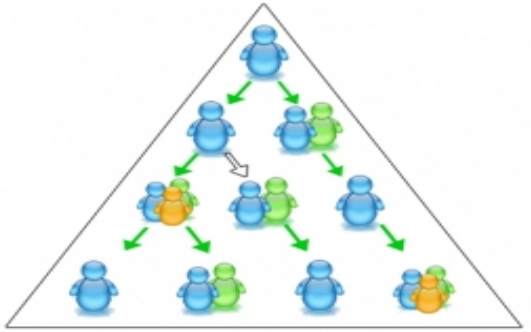
How frequently do you share recommendations online?



Viral Marketing and Influence Maximization

Business goal (Viral Marketing): exploit the "word-of-mouth" effect in a social network to achieve marketing objectives through self-replicating viral processes

Mining problem: find a seed-set of influential people such that by targeting them we maximize the spread of viral propagations



Hot results presented in:

"Mining the network value of customers" (Domingos et al, 2001)

"Mining knowledge-sharing sites for viral marketing" (Domingos et al, 2002)

"Maximizing the spread of influence through a social network" (Kempe et al, 2003)

The AIR Propagation Model

Authoritativeness of a user in a topic: $A_{u,z}$

Interest of a user for a topic: $S_{u,z}$

Relevance of an item for a topic: $\Phi_{i,z}$

Lady Gaga @ladygaga

Justin Bieber @justinbieber

Barack Obama @barackobama

CNN
@cnn

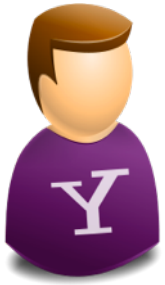
The Economist
@TheEconomist

0.92 0.08

0.01 0.09

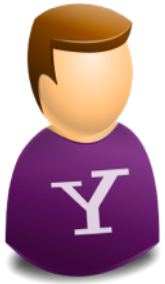
Presented in: Topic-aware social influence propagation models (Manco et al, 2012)

Explaining user behavior: Socio-topical relationships



- ✓ **Has good friends in Barcelona**
- ✓ **Does research on web mining**
- ✓ **Likes blues music**

Modeling socio-topical relationships



- ✓ **Has good friends in Barcelona**
- ✓ **Does research on web mining**
- ✓ **Likes blues music**

Who to follow - Refresh

Nicola Barbieri

Follow

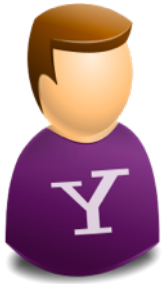
Francesco Bonchi

Follow

Giuseppe Manco

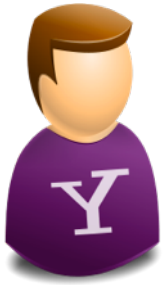
Follow

Modeling socio-topical relationships



- ✓ **Has good friends in Barcelona**
- ✓ **Does research on web mining**
- ✓ **Likes blues music**

Modeling socio-topical relationships



- ✓ **Has good friends in Barcelona**
- ✓ **Does research on web mining**
- ✓ **Likes blues music**

Who to follow - Refresh

Nicola Barbieri

Friend with **@ax**, **@bz**, **@bcn_fun**.

Follow

Francesco Bonchi

Authoritative about **#YahooLabs**, **#ViralMarketing**,**#WebMining**.

Follow

Giuseppe Manco

Authoritative about **#ClassicRock**, **#Blues**, **#AcousticGuitar**.

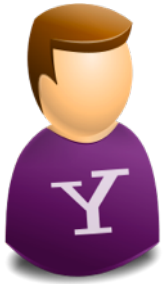
Follow

Common identity and common bond theory:

- Identity-based attachment holds when people join a community based on their interest in a well-defined common topic;
- Bond-based attachment is driven by personal social relations with other specific individuals.

Modeling socio-topical relationships

Identity Bond



- ✓ **Has good friends in Barcelona**
- ✓ **Does research on web mining**
- ✓ **Likes blues music**

Common identity and common bond theory:

- Identity-based attachment holds when people join a community based on their interest in a well-defined common topic;
- Bond-based attachment is driven by personal social relations with other specific individuals.

Presented in: Who to follow and why: link prediction with explanations (Barbieri et al, 2014)

Product Adoption Maximization

- Classical diffusion models such as Independet Cascade and Linear Threshold do not distinguish between influence and product adoption
- They implicitly assume that once influenced, a node necessarily adopts a product and that adopters always influence other users to adopt the product
- Sometimes influenced users, once they become active, may choose to not adopt but instead tattle about the product; by doing so, they may either promote or inhibit adoption by other users
- A propagation model called LT-C model that accounts for these observations can be used to investigate product adoption

Presented in: Maximizing Product Adoption in Social Networks (Lakshmanan et al, 2012)

Incentivized Social Advertising

- It allows influential user of a SN to get some money on the advertising revenue
- A study on incentivized social advertising formulate the problem of revenue maximization from the host perspective, when the incentives paid to the seed users are determined by their demonstrated past influence in the topic of the specific ad
- Two greedy algorithms for the problem:
  - CA-GREEDY is agnostic to users' incentives during the seed selection
  - CS-GREEDY is not

Presented in: Revenue Maximization in Incentivized Social Advertising (Aslay et al, 2017)

BAD CAMPAIGN VS GOOD CAMPAIGN

- Scenario: competing campaigns in a social network (Multi-Campaign Independent Cascade (MCICM))
- Problem: influence limitation where a “bad" campaign starts propagating from a certain node in the network
- Solution: use the notion of limiting campaigns to counteract the effect of misinformation
  - Performed by identifying a subset of individuals that need to be convinced to adopt the competing (or “good") campaign so as to minimize the number of people that adopt the “bad" campaign at the end of both propagation processes

Presented in: Limiting the Spread of Misinformation in Social Networks (Agrawal et al, 2011)

VIRAL MARKETING FROM HOST PERSPECTIVE

- Scenario: two or more players compete with similar products on the same network (*competitive viral marketing)*
- Problem: From the host's perspective, it is important not only to choose the seeds to maximize the collective expected spread, but also to assign seeds to companies so that it guarantees the "bang for the buck" for all companies is nearly identical
- Solution: *Needy Greedy* a propagation model capturing the competitive nature of viral marketing

Presented in: The Bang for the Buck: Fair Competitive Viral Marketing from the Host (Lu et al, 2013)

NON PROGRESSIVE MODELING

- Scenario: a user of a social network may stop using an app and become inactive, but again activate when instigated by a friend, or when the app adds a new feature or releases a new version
- Problem: The progressive model for influence maximization is no more valid
- Solution: Influence propagation can be modeled as a continuous-time Markov process with 2 states: active and inactive and compute the current state accordingly

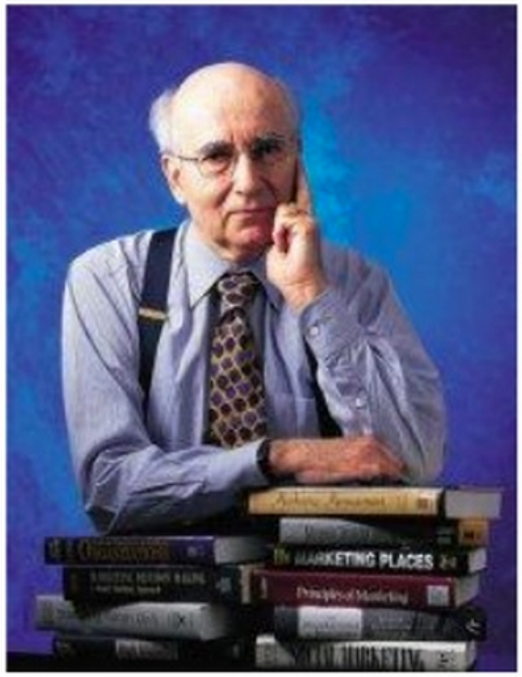
Presented in: Modeling Non-Progressive Phenomena for Influence Propagation (Lou et al, 2014)

TAKING INTO ACCOUNT CROWD DIVERSITY

- Scenario: How to consider the magnitude of influence and the diversity of the influenced crowd simultaneously
- Problem: Construct a class of diversity measures to quantify the diversity of the influenced crowd
- Solution: Formulate it as an optimization problem, i.e., *diversified social influence maximization*

Presented in: Diversified Social Influence Maximization (Tang et al, 2014)

Measuring Human Behavior In OSN

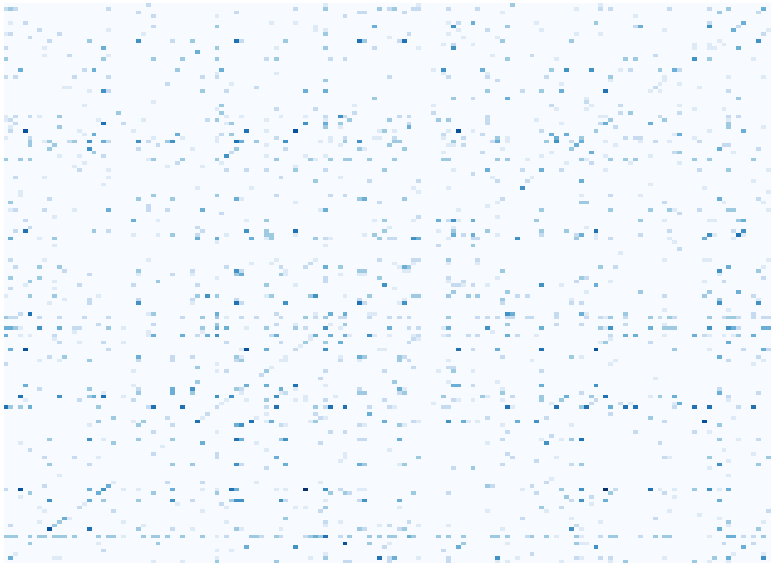


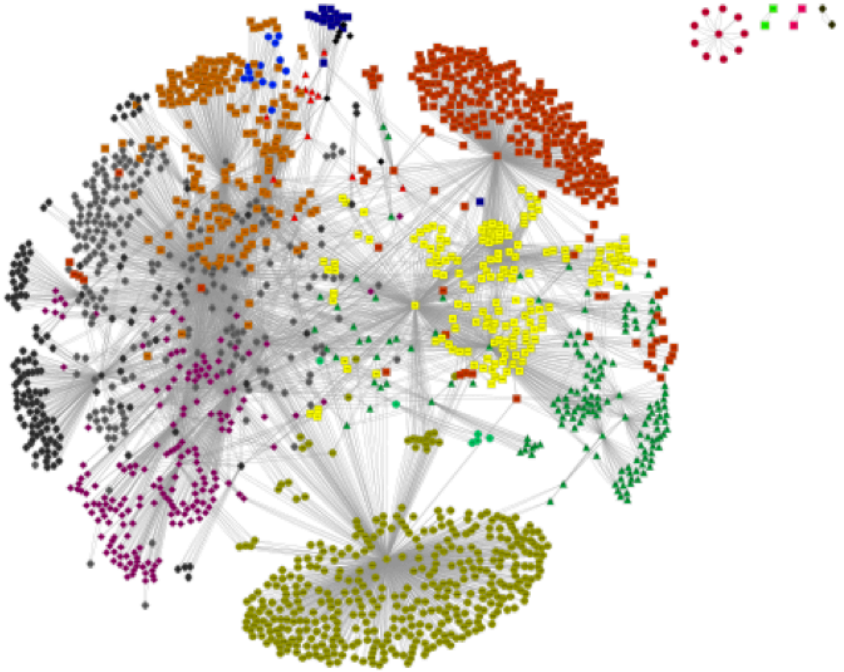
"Behavioral segmentation divides buyers into groups based on their knowledge, attitudes, uses or responses to a product."

Principles of Marketing, Kotler, Armstrong (2005)

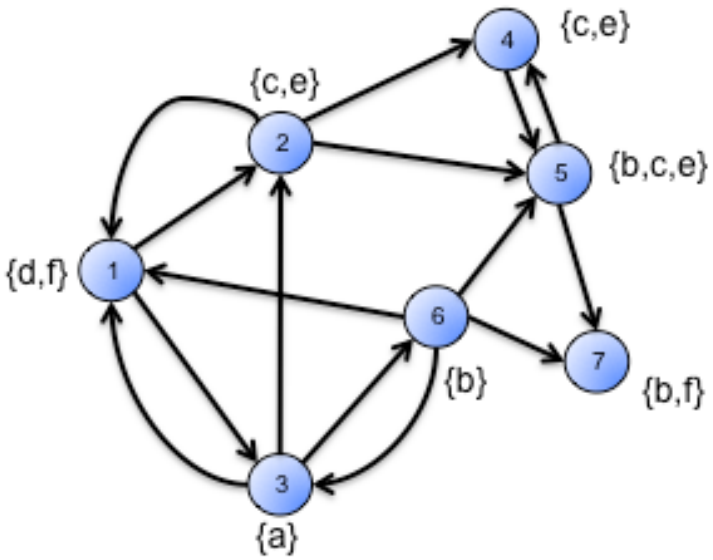
Measuring Human Behavior

- Tools for Processing contents, aimed at identifying
  - Groups and members
  - Topics and sentiments





Modeling social relationships and adoptions

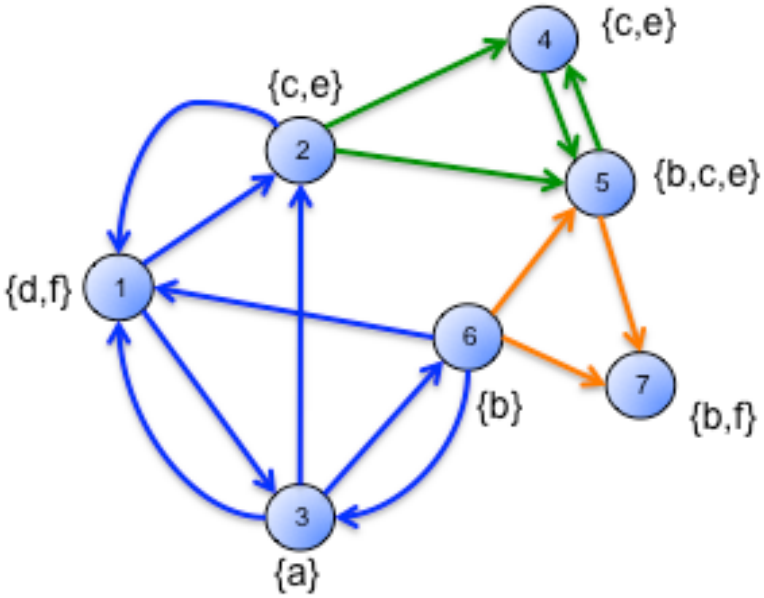


- Directed attributed-graph
- {1,2,3,4,5,6,7} user-set
- Links encode following relationships
- {a,b,c,d,e,f} features adopted by users

  E.g. hashtags, tags, products purchased

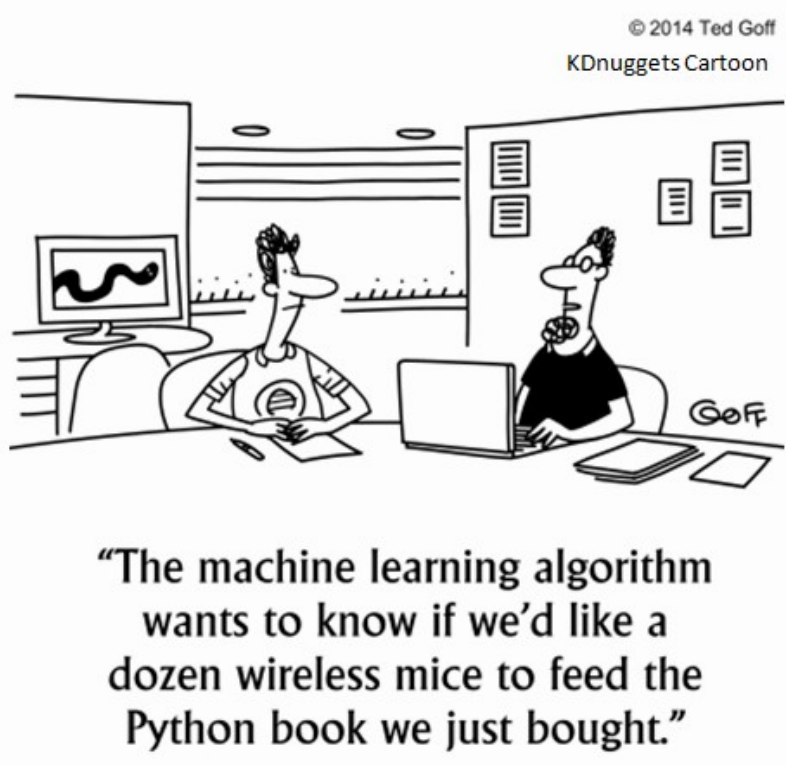
Credits: Giuseppe Manco, ICAR-CNR

An example: communities of like-minded people

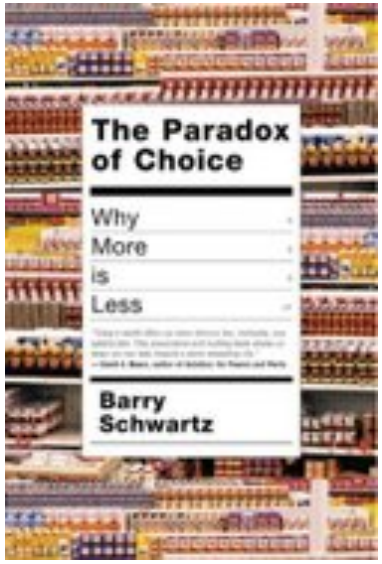


- **3 communities:**
  - **Blue links are bond-based;**
  - **Green and orange links are identity-based.**
- **Bond-based communities tend to have high density and reciprocal links**
- **Identity-based communities tend to exhibit a clear directionality**

Recommender Systems



Recommender Systems



250 varieties of cookies, 75 iced teas, 230 soups, 175 salad dressings, 275 cereals and 40 toothpastes

As the number of choices increases, it also increases the likelihood that we will make the WRONG one

Recommender Systems

"We are entering the age of recommendations"

[Henrik Schinzel,Avail Intelligence]

Recommender Systems

Recommender Systems are reshaping the world of e-commerce, helping customers find and purchase products, such as songs, books, movies, or news

The aim is to **transform a regular user into a buyer**

As the volumes of information grow, the importance of RS is likely to continue to grow and to have a key role in many **different industry domains**

Recommendation and knowledge discovery

Elisa

Recommendation and knowledge discovery

Analysis of user's profile

Analysis of preference data

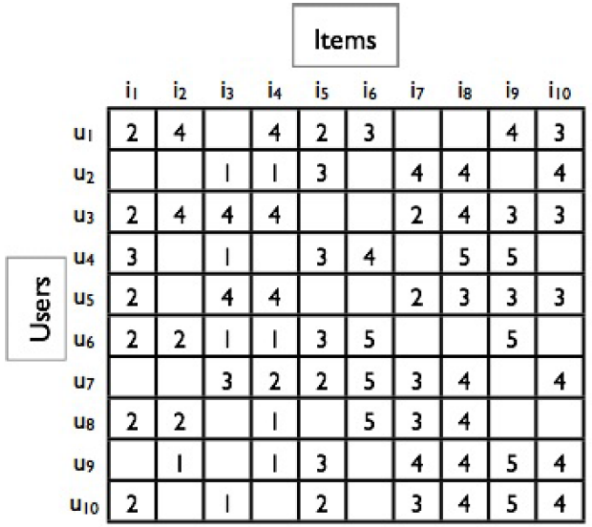
Analysis of user's preferences

Analysis of local trends and tendencies

Analysis of global trends and tendencies

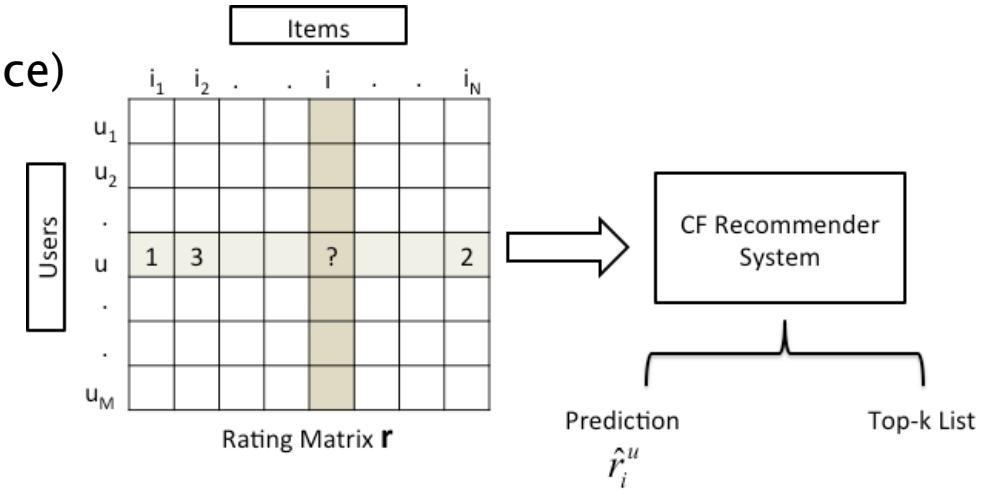
Preference data

- Implicit vs. explicit preferences
- Context
  - Time/seasonality
  - Dependency
  - Content
  - Demographic
  - Social relationship



Recommendation as matrix completion

- Prediction problem
  - predict the value of the missing entries
  - Given (u,i) predict value r
  - Use the predicted value to build a recommendation list
- Not just recommendation
  - (user, movie) (collaborative filtering)
  - (user, user) (link prediction)
  - … (item response, political science)



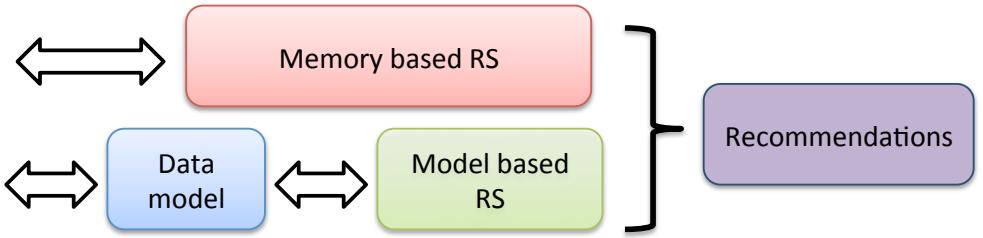
Evaluating Recommendations

- One of the biggest challenges in the area
- Offline evaluation: RMSE, Prec/Rec, AUC on test data
- Online evaluation: A/B test, empirical ROI

Approaches to recommendation

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	2	5			3	1		2		
u_2			2		4		5		1	
u_3	1		4			5		1		
u_4		4			4		2		2	
u_5	1		3		3	3		5		
u_6				2	1	3			4	5
u_7	4			5	5			1	1	
u_8	3	4		3			1		3	
u_9	1				3	2			4	4
u_{10}		5				5			5	4

Users' preference matrix



- Nearest Neighbor model
  - Given *(u,i)*, search for similar users/items
    - Need for a similarity $s_{u,v}$ ($s_{i,j}$)
  - Use neighbor information to predict

$$\hat{r}_i^u = \frac{\sum_{v \in \mathcal{N}^K(u)} s_{u,v} \cdot r_i^v}{\sum_{v \in \mathcal{N}^K(u)} s_{u,v}}$$

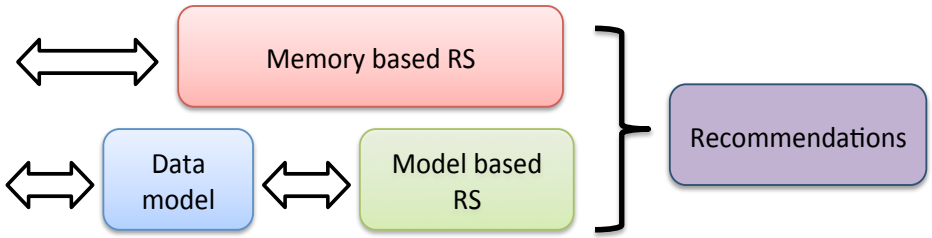
- Model-based approaches
  - Learn latent features $U_{[M \times K]}$ e $V_{[K \times N]}$ representing a decomposition of the original preference matrix
  - Exploit latent features to predict

$$\hat{r}_i^u = \sum_{k=1}^{K} \mathbf{U}_{u,k} \cdot \mathbf{V}_{k,i}.$$

Approaches to recommendation

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	2	5			3	1		2		
u_2			2		4		5		1	
u_3	1		4			5		1		
u_4		4			4		2		2	
u_5	1		3		3	3		5		
u_6				2	1	3			4	5
u_7	4			5	5			1	1	
u_8	3	4		3			1		3	
u_9	1				3	2			4	4
u_{10}		5				5			5	4

Users' preference matrix



- Neighbor models
  - Not scalable
  - Sensitive to the similarity function
- Latent factor models
  - U and V not easy to interpret
  - Prone to overfitting
  - Low RMSE, low Prec/Rec
    - However, learning can be tuned to different loss functions

Probabilistic modeling

- Treat data as observations that arise from a generative probabilistic process that includes hidden variables
  - For preference data, the hidden variables reflect the behavior of single users or the commonality in items
- Infer the hidden structure using posterior inference
  - What are the latent behaviors that describe this group of users?
- Situate new data into the estimated model.
  - How does a new user fit into the estimated behavioral structure?

Mixture Membership models for recommendation

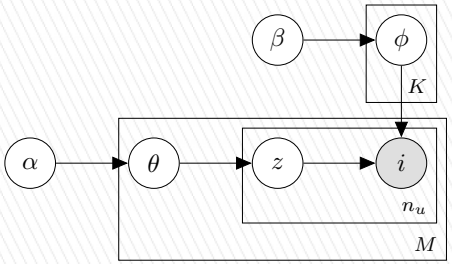
Recommending items with Mixture Membership models

- Prospective items are ranked exploiting the components

$$r_i^u \propto \sum_k \vartheta_{u,k} \varphi_{k,i}$$

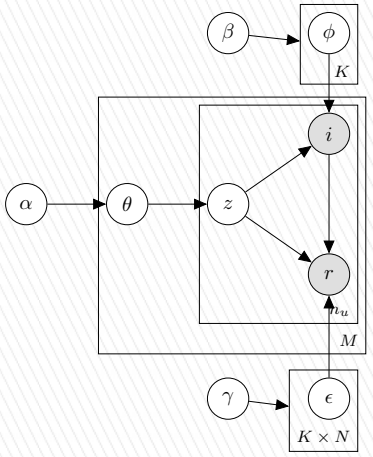
- Top-ranked items enter a recommendation list
  - LDA ranking performs significantly better than PMF, SVD,…
- Several variations of the basic model
  - E.g. Bayesian User Community considers both ratings and implicit selection

Graphically



1. For each latent factor $k = 1, \ldots, K$ sample a multinomial distribution $\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta})$
2. For each user $u \in \mathcal{U}$
   (a) Sample the number n_u of item selections;
   (b) Choose $\boldsymbol{\theta}_u \sim Dir(\boldsymbol{\alpha})$
   (c) For each of the n_u items to be generated:
      i. Sample a topic $z \sim Disc(\boldsymbol{\theta}_u)$
      ii. Sample $i \sim Disc(\boldsymbol{\phi}_z)$

LDA



1. For each latent factor $z \in \{1, \cdots, K\}$,
   (a) Sample item selection components $\boldsymbol{\phi}_z \sim Dir(\boldsymbol{\beta})$
   (b) for each item $i \in \mathcal{I}$ sample rating probabilities $\boldsymbol{\varepsilon}_{z,i} \sim Dir(\boldsymbol{\gamma})$
2. For each user $u \in \mathcal{U}$
   (a) sample user community-mixture components $\boldsymbol{\theta}_u \sim Dir(\boldsymbol{\alpha})$
   (b) Sample the number of items n_u for the user u
   (c) For each of the n_u items to select
      i. sample a latent factor $z \sim Disc(\boldsymbol{\theta}_u)$
      ii. Choose an item $i \sim Disc(\boldsymbol{\phi}_z)$
      iii. Generate a rating value $r \sim Disc(\boldsymbol{\varepsilon}_{z,i})$

Bayesian User Community Model

- **An Analysis of Probabilistic Methods for Top-N Recommendation in Collaborative Filtering (Manco et al, 2011)**
- **Modeling item selection and relevance for accurate recommendations: a bayesian approach (Costa et al, 2011)**
- **A Probabilistic Hierarchical Approach for Pattern Discovery in Collaborative Filtering Data (Ritacco et al, 2011)**

Outline

- A brief history of Social Networks (SN)
- The Big Data Challenges
- Social Networks (SN) Big Data Features
- What happened so far
- **Conclusions**

Conclusions

- SN and Big Data poses so many challenges…
- Many challenges = Many opportunities for research and industry
- It is necessary a multidisciplinary approach that integrates technical and sociological skills
- Think out of the box but study all the old school statistics and sociology…

THANK YOU