Introduction and Overview

APAM E4990
Modeling Social Data

Jake Hofman

Columbia University

January 20, 2017

Course overview

Modeling social data requires an understanding of:

1. How to obtain data produced by (online) human interactions,
2. What questions we typically ask about human-generated data,
3. How to reframe these questions as mathematical models, and
4. How to interpret the results of these models in ways that address our questions.

Questions

Many long-standing questions in the social sciences are notoriously difficult to answer, e.g.:

- "Who says what to whom in what channel with what effect"? (Laswell, 1948)
- How do ideas and technology spread through cultures? (Rogers, 1962)
- How do new forms of communication affect society? (Singer, 1970)
- ...

Questions

Typically difficult to observe the relevant information via conventional methods

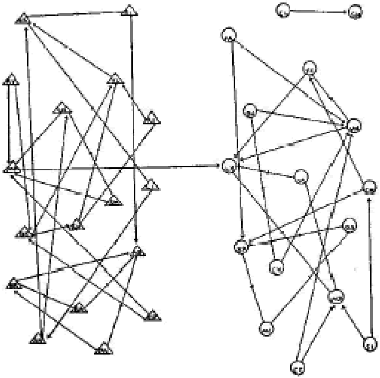

EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

FIRST STUDIES EXHIBITED

Colored Lines Show Likes and Dislikes of Individuals and of Groups.

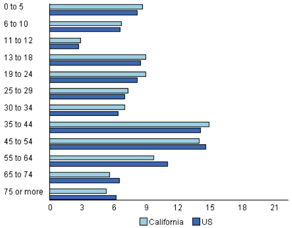
MANY MISFITS REVEALED



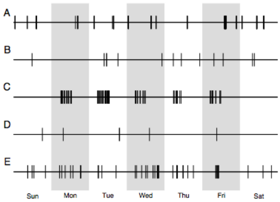
Moreno, 1933

Large-scale data

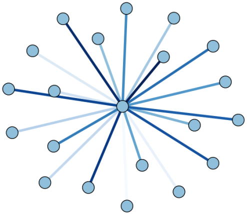
Recently available electronic data provide an unprecedented opportunity to address these questions at scale



Demographic



Behavioral



Network

Computational social science

An emerging discipline at the intersection of the social sciences, statistics, and computer science

Computational social science

An emerging discipline at the intersection of the social sciences, statistics, and computer science

(motivating questions)

Computational social science

An emerging discipline at the intersection of the social sciences, statistics, and computer science

(fitting large, potentially sparse models)

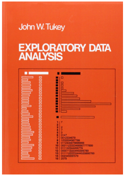
Computational social science

An emerging discipline at the intersection of the social sciences, statistics, and computer science

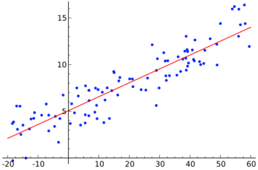
(parallel processing for filtering and aggregating data)

Topics

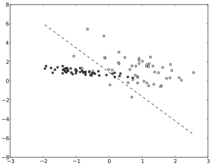
Exploratory Data Analysis



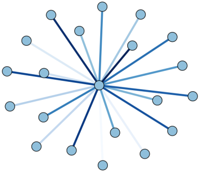
Regression



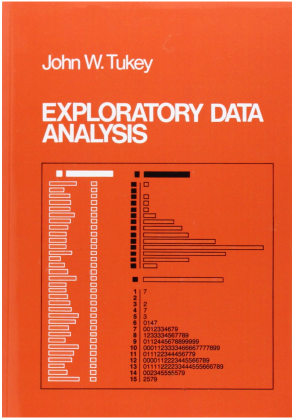
Classification



Networks

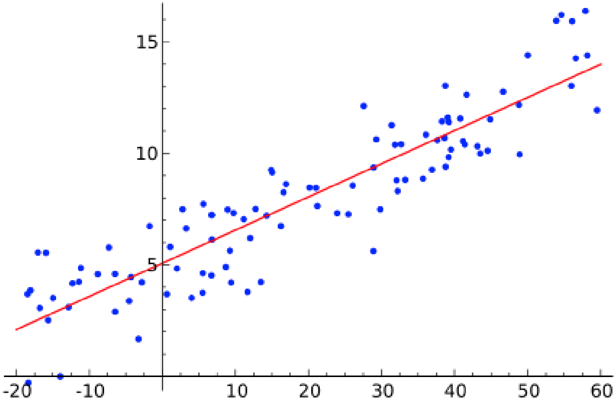


Exploratory Data Analysis



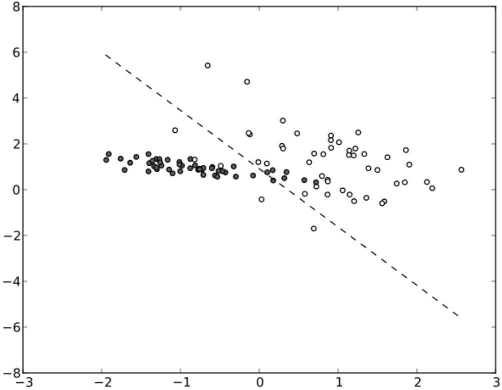
(a.k.a. counting and plotting things)

Regression



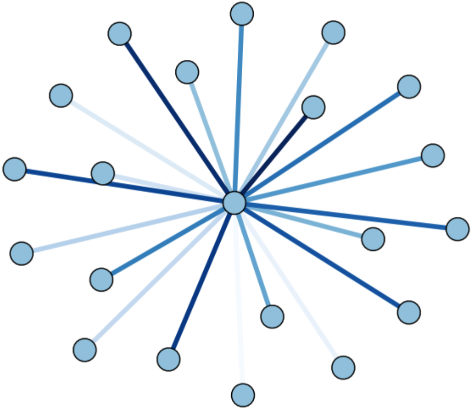
(a.k.a. modeling continuous things)

Classification



(a.k.a. modeling discrete things)

Networks



(a.k.a. counting complicated things)

Topics

Date	Topic
2017-01-20	Introduction: Case Studies
2017-01-27	Counting: Split/Apply/Combine
2017-02-03	Counting at Scale: MapReduce
2017-02-10	Computational complexity
2017-02-17	Data visualization
2017-02-24	Regression I: Theory and Practice
2017-03-03	Regression II: Theory and Practice
2017-03-10	Classification I: Naive Bayes
2017-03-17	*Spring Break*
2017-03-24	Classification II: Logistic Regression
2017-03-31	Networks I: Representations, characteristics
2017-04-07	Networks II: Counting on graphs
2017-04-14	Causality and Experiments: II
2017-04-21	Causality and Experiments: II
2017-04-28	Student Presentations

http://modelingsocialdata.org

The ~~clean~~ real story

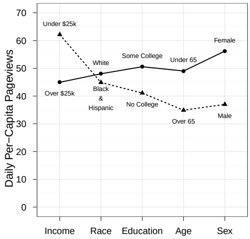
*"**We have a habit** in writing articles published in scientific journals to make the work as finished as possible, **to cover all the tracks**, to not worry about the blind alleys or to describe how you had the wrong idea first, and so on. So there **isn't any place to publish**, in a dignified manner, **what you actually did** in order to get to do the work ..."*

-Richard Feynman
Nobel Lecture[1], 1965

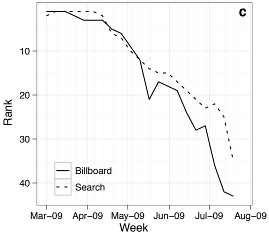
[1] http://bit.ly/feynmannobel

Case studies

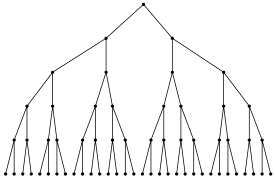
Web demographics



Search predictions

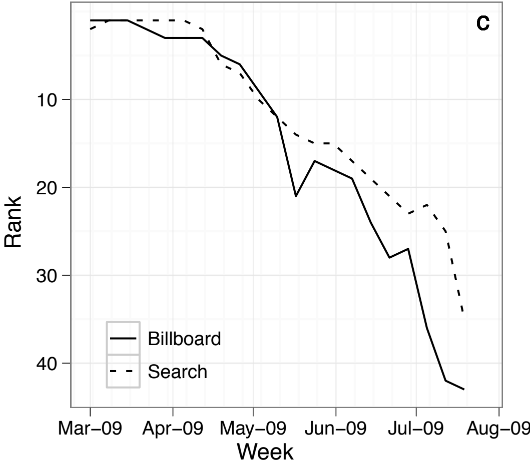


Viral hits



Predicting consumer activity with Web search

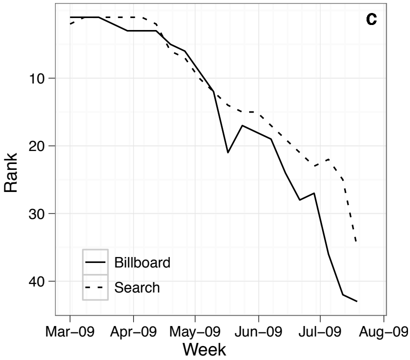
with Sharad Goel, Sébastien Lahaie, David Pennock, Duncan Watts



Search predictions

Motivation

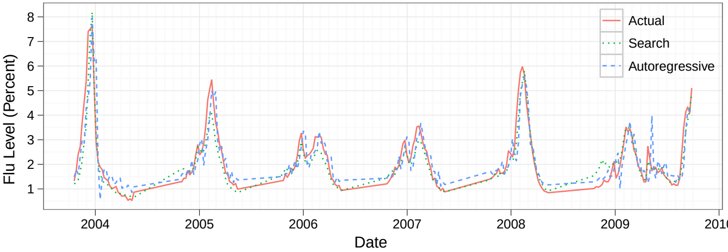
Does collective search activity provide useful predictive signal about real-world outcomes?



Search predictions

Motivation

Past work mainly focuses on predicting the present[2] and ignores baseline models trained on publicly available data

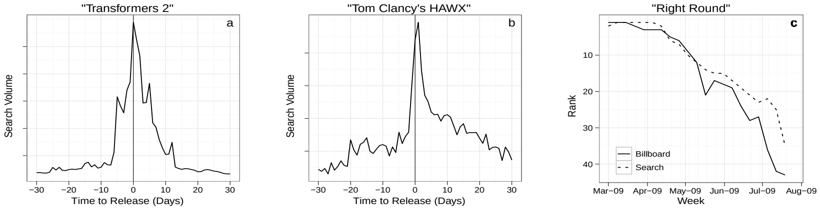


[2] Varian, 2009

Search predictions

Motivation

We predict future sales for movies, video games, and music



Search predictions

Search models

For movies and video games, predict opening weekend box office and first month sales, respectively:

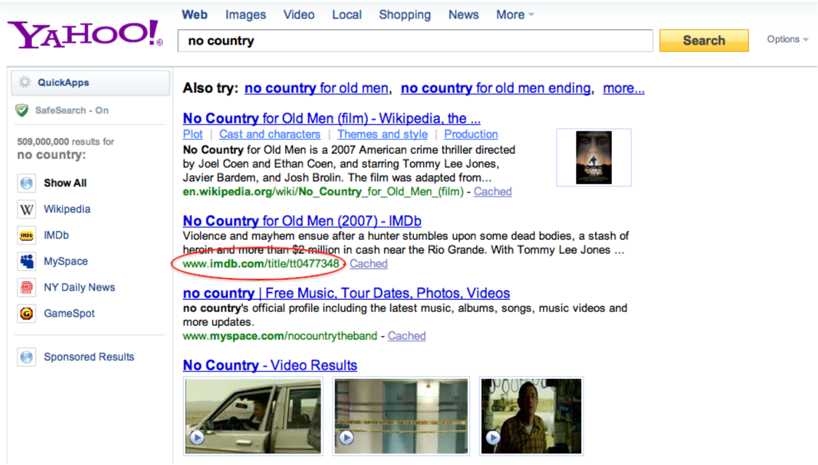
$$\log(\mathbf{revenue}) = \beta_0 + \beta_1 \log(\mathbf{search}) + \epsilon$$

For music, predict following week's Billboard Hot 100 rank:

$$\mathbf{billboard}_{t+1} = \beta_0 + \beta_1 \mathbf{search}_t + \beta_2 \mathbf{search}_{t-1} + \epsilon$$

Search predictions

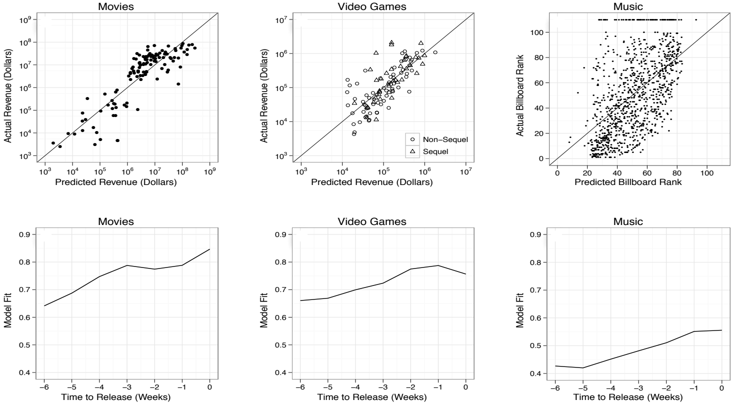
Search volume



Search predictions

Search models

Search activity is predictive for movies, video games, and music weeks to months in advance



Search predictions

Baseline models

For movies, use budget, number of opening screens and Hollywood Stock Exchange:

$$\log(\textbf{revenue}) = \beta_0 + \beta_1 \log(\textbf{budget}) + \beta_2 \log(\textbf{screens}) + \beta_3 \log(\textbf{hsx}) + \epsilon$$

Search predictions

Baseline models

For video games, use critic ratings and predecessor sales (sequels only):

$$\log(\textbf{revenue}) = \beta_0 + \beta_1 \textbf{rating} + \beta_2 \log(\textbf{predecessor}) + \epsilon$$

Search predictions

Baseline models

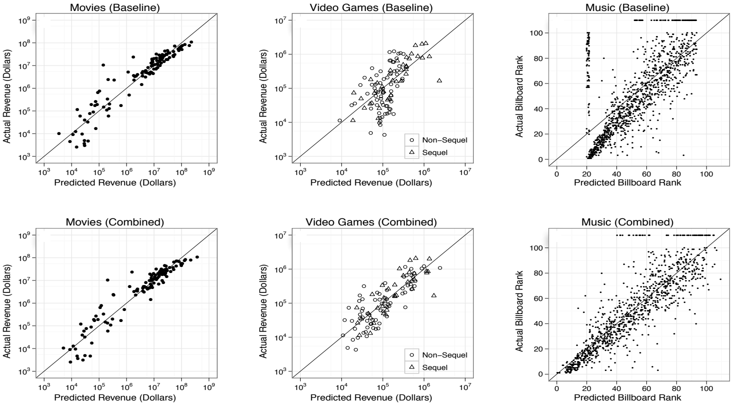
For music, use an autoregressive model with the previously available rank:

$$\textbf{billboard}_{t+1} = \beta_0 + \beta_1 \textbf{billboard}_{t-1} + \epsilon$$

Search predictions

Baseline + combined models

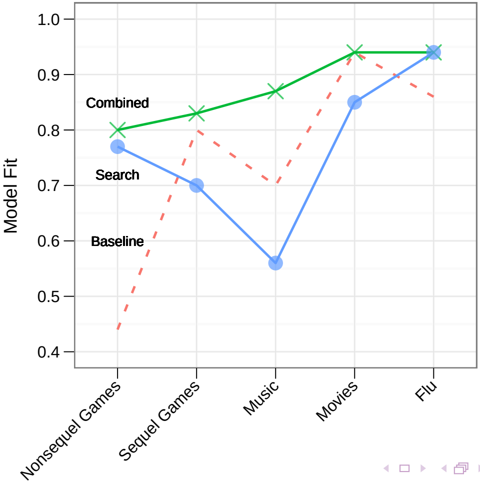
Baseline models are often surprisingly good



Search predictions

Model comparison

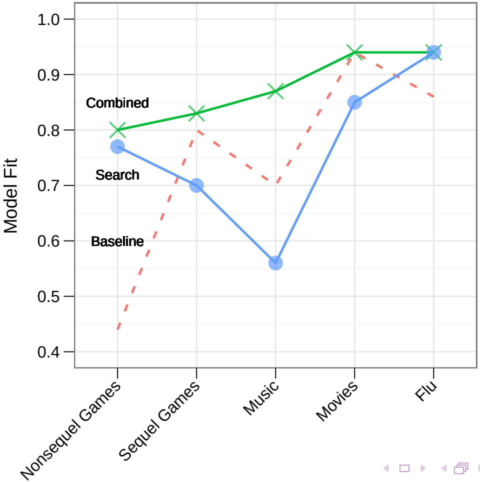
For movies, search is outperformed by the baseline and of little marginal value



Search predictions

Model comparison

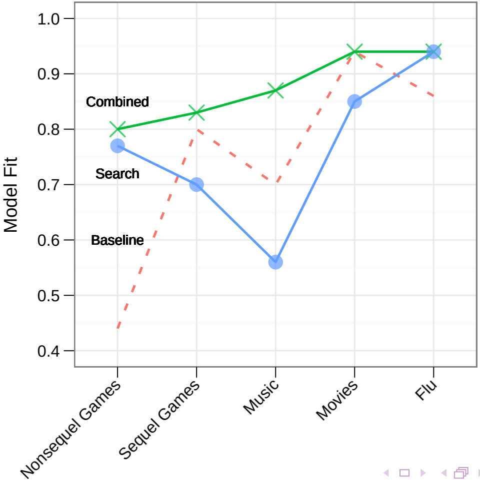
For video games, search helps substantially for non-sequels, less so for sequels



Search predictions

Model comparison

For music, the addition of search yields a substantially better combined model



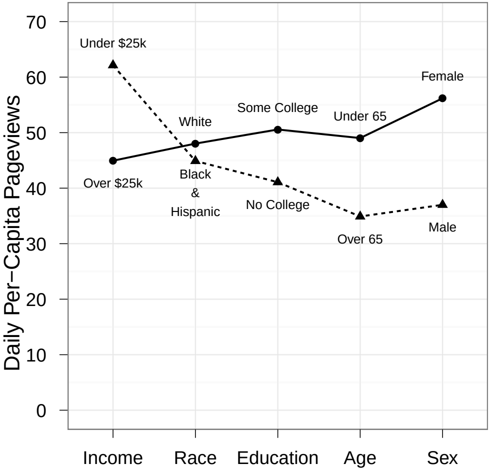
Search predictions

Summary

- Relative performance and value of search varies across domains
- Search provides a fast, convenient, and flexible signal across domains
- "Predicting consumer activity with Web search"
  Goel, Hofman, Lahaie, Pennock & Watts, PNAS 2010

Demographic diversity on the Web

with Irmak Sirer and Sharad Goel (ICWSM 2012)



Motivation

Science 17 April 1998:
Vol. 280 no. 5362 pp. 390-391
DOI: 10.1126/science.280.5362.390

‹ Prev | Table of Contents | Next ›

POLICY

INFORMATION ACCESS

Bridging the Racial Divide on the Internet

Donna L. Hoffman and Thomas P. Novak

+ Author Affiliations

The Internet is expected to do no less than transform society (1); its use has been increasing exponentially since 1994 (2). But are all members of our society equally likely to have access to the Internet and thus participate in the rewards of this transformation? Here we present findings both obvious and surprising from a recent survey of Internet access and discuss their implications for social science research and public policy.

Previous work is largely survey-based and focuses and group-level differences in online access

Motivation

"As of January 1997, we estimate that 5.2 million African Americans and 40.8 million whites have ever used the Web, and that 1.4 million African Americans and 20.3 million whites used the Web in the past week."

-Hoffman & Novak (1998)

Motivation

Focus on activity instead of access

How diverse is the Web?

To what extent do online experiences vary across demographic groups?

Data

nielsen MegaPanel

- Representative sample of 265,000 individuals in the US, paid via the Nielsen MegaPanel[3]
- Log of anonymized, complete browsing activity from June 2009 through May 2010 (URLs viewed, timestamps, etc.)
- Detailed individual and household demographic information (age, education, income, race, sex, etc.)

[3] Special thanks to Mainak Mazumdar

Data

```
# ls -alh nielsen_megapanel.tar
-rw-r--r--  100G Jul 17 13:00 nielsen_megapanel.tar
```

Data

```
# ls -alh nielsen_megapanel.tar
-rw-r--r--  100G Jul 17 13:00 nielsen_megapanel.tar
```

- Normalize pageviews to at most three domain levels, sans `www`
  e.g. `www.yahoo.com` → `yahoo.com`,
  `us.mg2.mail.yahoo.com/neo/launch` → `mail.yahoo.com`

Data

```
# ls -alh nielsen_megapanel.tar
-rw-r--r--  100G Jul 17 13:00 nielsen_megapanel.tar
```

- Normalize pageviews to at most three domain levels, sans `www`
  e.g. `www.yahoo.com` → `yahoo.com`,
  `us.mg2.mail.yahoo.com/neo/launch` → `mail.yahoo.com`
- Restrict to top 100k (out of 9M+ total) most popular sites (by unique visitors)

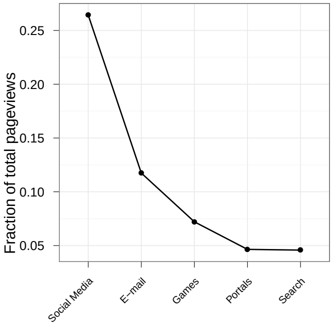
Data

```
# ls -alh nielsen_megapanel.tar
-rw-r--r--  100G Jul 17 13:00 nielsen_megapanel.tar
```

- Normalize pageviews to at most three domain levels, sans `www`
  e.g. `www.yahoo.com` → `yahoo.com`,
  `us.mg2.mail.yahoo.com/neo/launch` → `mail.yahoo.com`
- Restrict to top 100k (out of 9M+ total) most popular sites (by unique visitors)
- Aggregate activity at the site, group, and user levels

Aggregate usage patterns

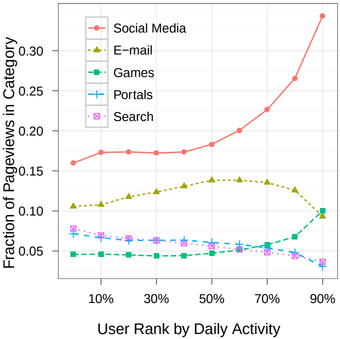
How do users distribute their time across different categories?



All groups spend the majority of their time in the top five most popular categories

Aggregate usage patterns

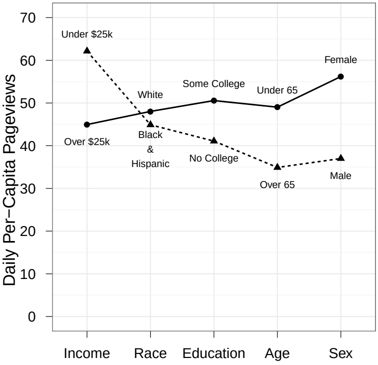
How do users distribute their time across different categories?



Highly active users devote nearly twice as much of their time to social media relative to typical individuals

Group-level activity

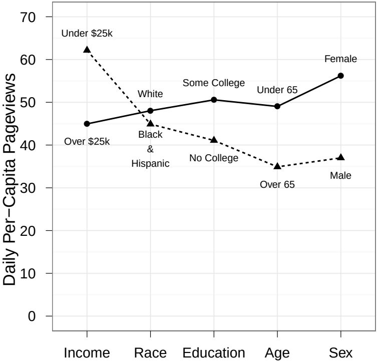
How does browsing activity vary at the group level?



Large differences exist even at the aggregate level
(e.g. women on average generate 40% more pageviews than men)

Group-level activity

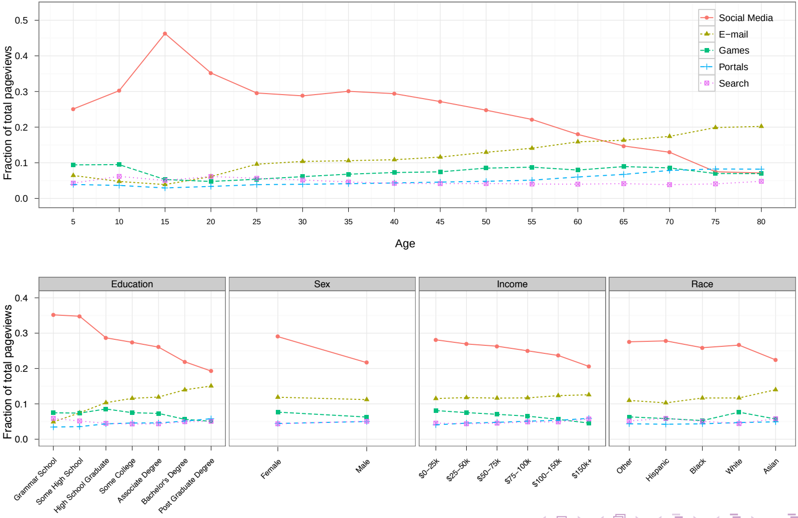
How does browsing activity vary at the group level?



Younger and more educated individuals are both more likely to access the Web and more active once they do

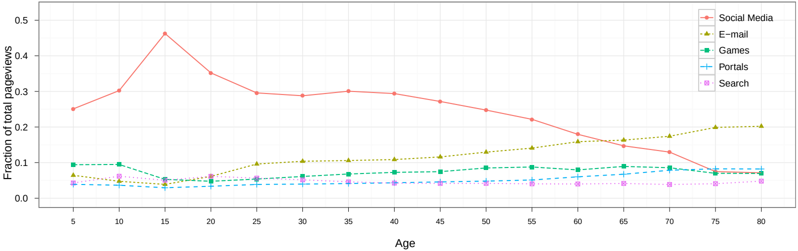
Group-level activity

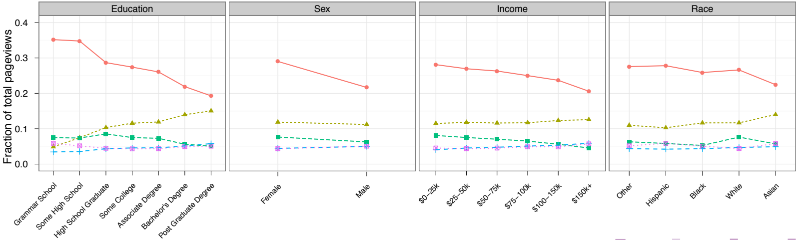
All demographic groups spend the majority of their time in the same categories



Group-level activity

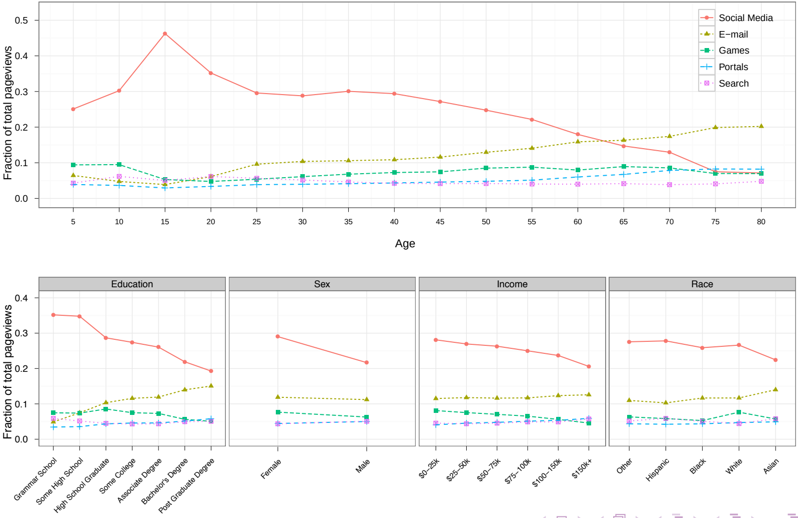
Older, more educated, male, wealthier, and Asian Internet users spend a smaller fraction of their time on social media





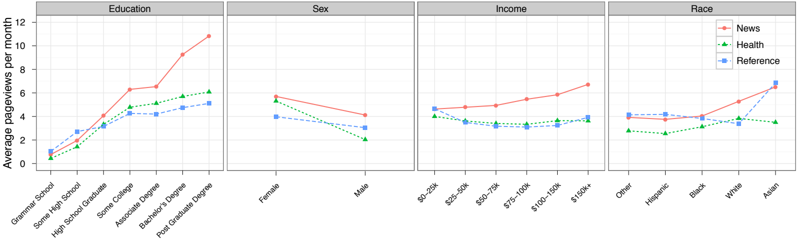
Group-level activity

Lower social media use by these groups is often accompanied by higher e-mail volume



Revisiting the digital divide

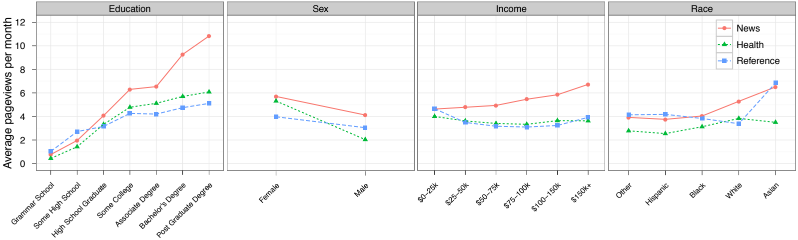
How does usage of news, health, and reference vary with demographics?



Post-graduates spend three times as much time on health sites than adults with only some high school education

Revisiting the digital divide

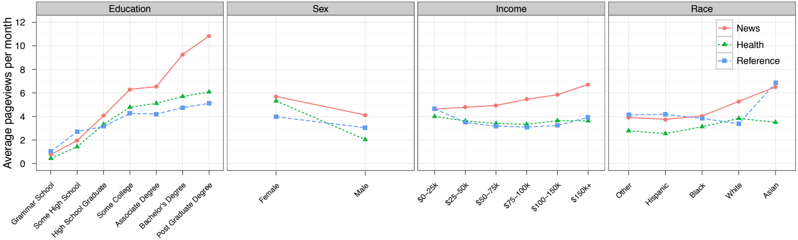
How does usage of news, health, and reference vary with demographics?



Asians spend more than 50% more time browsing online news than do other race groups

Revisiting the digital divide

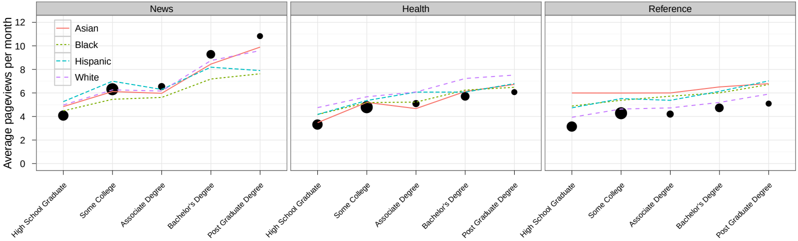
How does usage of news, health, and reference vary with demographics?



Even when less educated and less wealthy groups gain access to the Web, they utilize these resources relatively infrequently

Revisiting the digital divide

How does usage of news, health, and reference vary with demographics?

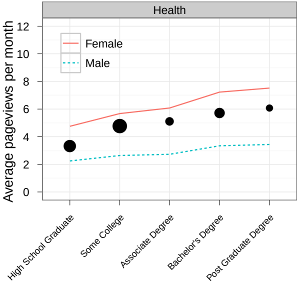


Controlling for other variables, effects of race and gender largely disappear, while education continues to have large effect

$$p_i = \sum_j \alpha_j x_{ij} + \sum_j \sum_k \beta_{jk} x_{ij} x_{ik} + \sum_j \gamma_j x_{ij}^2 + \epsilon_i$$

Revisiting the digital divide

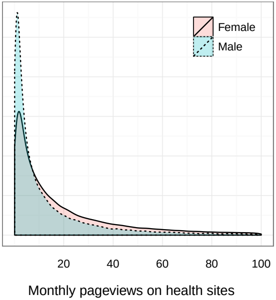
How does usage of news, health, and reference vary with demographics?



However, women spend considerably more time on health sites compared to men

Revisiting the digital divide

How does usage of news, health, and reference vary with demographics?



However, women spend considerably more time on health sites compared to men, although means can be misleading

Individual-level prediction

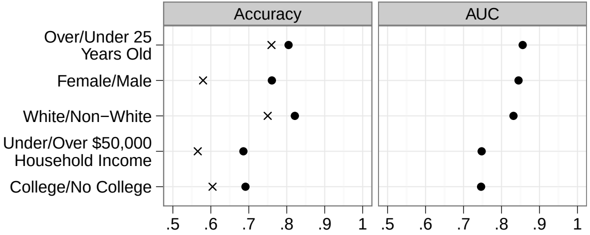
How well can one predict an individual's demographics from their browsing activity?

- Represent each user by the set of sites visited
- Fit linear models[4] to predict majority/minority for each attribute on 80% of users
- Tune model parameters using a 10% validation set
- Evaluate final performance on held-out 10% test set

[4] http://bit.ly/svmperf

Individual-level prediction

Reasonable (∼70-85%) accuracy and AUC across all attributes



Individual-level prediction

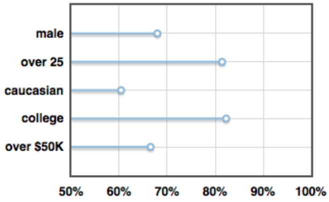
Highly-weighted sites under the fitted models

	Large positive weight	**Large negative weight**
Female	winster.com lancome-usa.com	sports.yahoo.com espn.go.com
White	marlboro.com cmt.com	mediatakeout.com bet.com
College Educated	news.yahoo.com linkedin.com	youtube.com myspace.com
Over 25 Years Old	evite.com classmates.com	addictinggames.com youtube.com
Household Income Under $50,000	eharmony.com tracfone.com	rownine.com matrixdirect.com

Individual-level prediction

Proof of concept browser demo

From the 28 sites we found in your browser history, it appears that you're a **caucasian male** who is **over 25** years old with a **college** education earning **over $50K** per year.



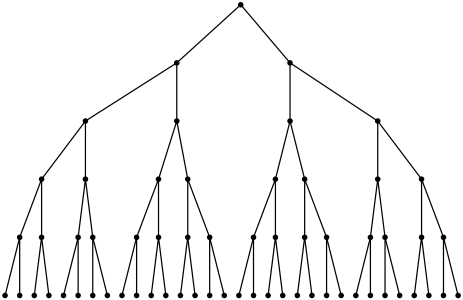
`http://bit.ly/surfpreds`

Summary

- Highly active users spend disproportionately more of their time on social media and less on e-mail relative to the overall population
- Access to research, news, and healthcare is strongly related to education, not as closely to ethnicity
- User demographics can be inferred from browsing activity with reasonable accuracy
- "Who Does What on the Web", Goel, Hofman & Sirer, ICWSM 2012

The structural virality of online diffusion

with Ashton Anderson, Sharad Goel, Duncan Watts (Management Science 2015)



"Going Viral"?

viral

Contents [show]

English

Etymology

From the stem of *virus* with suffix *-al*.

Pronunciation

- IPA: /ˈvaɪrəl/
- Rhymes: -aɪrəl

Adjective

viral (*not comparable*)

1. (*virology*) Of or relating to a biological virus.

   ***viral** DNA*

2. (*virology*) Caused by a virus.

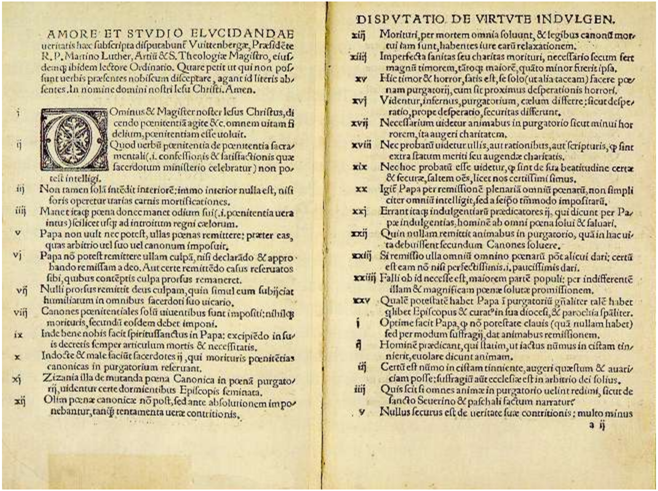
   ***viral** infection*

3. (*computing*) Of the nature of an informatic virus; able to spread copies of itself to other computers.
4. (*advertising and marketing*) Spread by word of mouth, with minimal intervention in order to create buzz and interest.

Derived terms

- go viral
- viral marketing

"Going Viral"?



AMORE ET STVDIO ELVCIDANDAE ueritatis hæc subscripta disputabunt Vuittenbergæ, Præsidēte R. P. Martino Luther, Artiū & S. Theologiæ Magistro, eiusdemq; ibidem lectore Ordinario. Quare petit ut qui non possunt uerbis præsentes nobiscum disceptare, agant id literis absentes. In nomine domini nostri Iesu Christi. Amen.

i Dominus & Magister noster Iesus Christus, dicendo pœnitentiā agite &c. omnem uitam fidelium, pœnitentiam esse uoluit.

ij Quod uerbū pœnitentia de pœnitentia sacramentali (.i. confessionis & satisfactionis quæ sacerdotum ministerio celebratur) non potest intelligi.

iij Non tamen solā intēdit interiorē; immo interior nulla est, nisi foris operetur uarias carnis mortificationes.

iiij Manet itaq; pœna donec manet odium sui (.i. pœnitentia uera intus) scilicet usq; ad introitum regni cælorum.

v Papa non uult nec potest, ullas pœnas remittere; præter eas, quas arbitrio uel suo uel canonum imposuit.

vj Papa nō potest remittere ullam culpā, nisi declarādo & approbando remissam a deo. Aut certe remittēdo casus reseruatos sibi, quibus contēptis culpa prorsus remaneret.

vij Nulli prorsus remittit deus culpam, quin simul eum subijciat humiliatum in omnibus sacerdoti suo uicario.

viij Canones pœnitentiales solū uiuentibus sunt impositi: nihilq; morituris, secundū eosdem debet imponi.

ix Inde bene nobis facit spiritussanctus in Papa: excipiēdo in suis decretis semper articulum mortis & necessitatis.

x Indocte & male faciūt sacerdotes ij, qui morituris pœnitētias canonicas in purgatorium reseruant.

xj Zizania illa de mutanda pœna Canonica in pœnā purgatorij, uidentur certe dormientibus Episcopis seminata.

xij Olim pœnæ canonicæ nō post, sed ante absolutionem imponebantur, tanq; tentamenta ueræ contritionis.

DISPVTATIO DE VIRTVTE INDVLGEN.

xiij Morituri, per mortem omnia soluunt, & legibus canonū mortui iam sunt, habentes iure earū relaxationem.

xiiij Imperfecta sanitas seu charitas morituri, necessario secum fert magnū timorem, tātoq; maiorē, quāto minor fuerit ipsa.

xv Hic timor & horror, satis est, se solo (ut alia taceam) facere pœnam purgatorij, cum sit proximus desperationis horrori.

xvj Videntur, infernus, purgatorium, cælum differre: sicut desperatio, prope desperatio, securitas differunt.

xvij Necessarium uidetur animabus in purgatorio sicut minui horrorem, ita augeri charitatem.

xviij Nec probatū uidetur ullis, aut rationibus, aut scripturis, q sint extra statum meriti seu augendæ charitatis.

xix Nec hoc probatū esse uidetur, q sint de sua beatitudine certæ & securæ, saltem oēs, licet nos certissimi simus.

xx Igit Papa per remissionē plenariā omniū pœnarū, non simpliciter omniū intelligit, sed a seipso tm̄modo impositarū.

xxj Errant itaq; indulgentiarū prædicatores ij, qui dicunt per Papæ indulgentias, hominē ab omni pœna solui & saluari.

xxij Quin nullam remittit animabus in purgatorio, quā in hac uita debuissent secundum Canones soluere.

xxiij Si remissio ulla omniū omnino pœnarū pōt alicui dari; certū est eam nō nisi perfectissimis .i. paucissimis dari.

xxiiij Falli ob id necesse est, maiorem partē populi: per indifferentē illam & magnificam pœnæ solutæ promissionem.

xxv Qualē potestatē habet Papa ī purgatoriū gn̄aliter talē habet qlibet Episcopus & curat' in sua diocesi, & parochia spāliter.

i Optime facit Papa, q nō potestate clauis (quā nullam habet) sed per modum suffragij, dat animabus remissionem.

ij Hominē prædicant, qui statim, ut iactus nūmus in cistam tinnierit, euolare dicunt animam.

iij Certū est nūmo in cistam tinniente, augeri quæstum & auariciam posse: suffragiū aūt ecclesiæ est in arbitrio dei solius.

iiij Quis scit si omnes animæ in purgatorio uelint redimi, sicut de sancto Seuerino & paschali factum narratur?

v Nullus securus est de ueritate suæ contritionis; multo minus

a ij

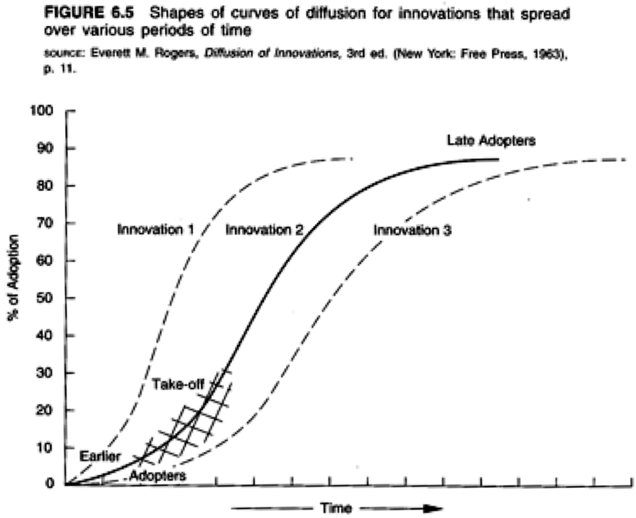
"Going Viral"?

"Therefore we … wish to proceed with great care as is proper, and to cut off the advance of this **plague** *and* **cancerous disease** *so it will not* **spread** *any further …"*[5]

-Pope Leo X
Exsurge Domine (1520)

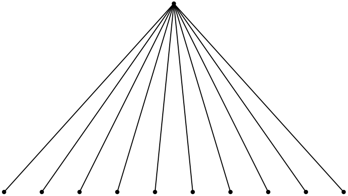
[5] http://www.economist.com/node/21541719

"Going Viral"?



Rogers (1962), Bass (1969)

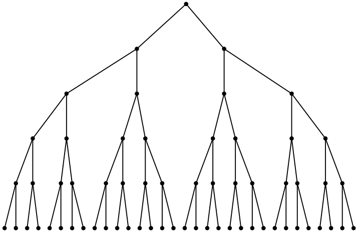
“Going viral”?




CNN Money
CNNMoney.com @CNNMoney
3 May
Dow crosses 15,000 for the first time, fueled by strong jobs report.
cnnmon.ie/bkgnews
Retweeted by CNN Breaking News
Collapse Reply Retweet Favorite More
273 RETWEETS
49 FAVORITES
10:23 AM - 3 May 13 · Details

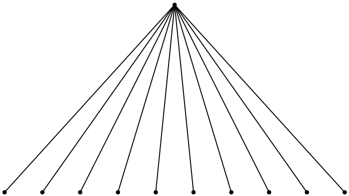
"Going viral"?

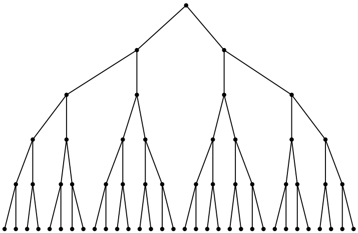

CWB Brasil queremos novamente o show da Banda Restart em Curitiba - Paraná
Created 12 months ago by @PeLuMoraComigo
Description
Pedimos atenciosamente a CWB Brasil novamente o show da Banda Restart em Curitiba. Desde o dia 29 de
2,352
signatures so far ...
Sign
Tweet my signature
#Twitition CWB Brasil queremos novamente o show da Banda Restart em Curitiba - Paraná http://twitition.com/opxpv
Follow @Twitition



"Going viral"?

How do popular things become popular?





Data

- Examined one year of tweets from July 2011 to July 2012

Data

- Examined one year of tweets from July 2011 to July 2012
- Restricted to 1.4 billion tweets containing links to top news, videos, images, and petitions sites

Data

- Examined one year of tweets from July 2011 to July 2012
- Restricted to 1.4 billion tweets containing links to top news, videos, images, and petitions sites
- Aggregated tweets by URL, resulting in 1 billion distinct "events"

Data

- Examined one year of tweets from July 2011 to July 2012
- Restricted to 1.4 billion tweets containing links to top news, videos, images, and petitions sites
- Aggregated tweets by URL, resulting in 1 billion distinct "events"
- Crawled friend list of each adopter

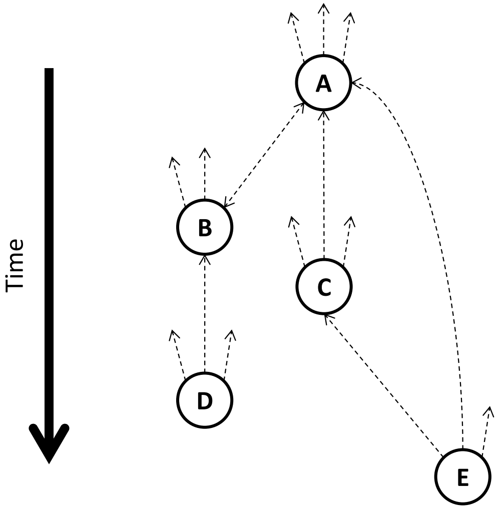
Data

- Examined one year of tweets from July 2011 to July 2012
- Restricted to 1.4 billion tweets containing links to top news, videos, images, and petitions sites
- Aggregated tweets by URL, resulting in 1 billion distinct "events"
- Crawled friend list of each adopter
- Inferred "who got what from whom" to construct diffusion trees

Data

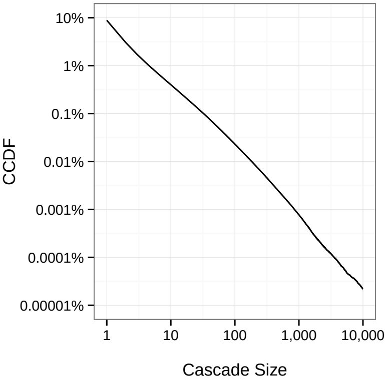
- Examined one year of tweets from July 2011 to July 2012
- Restricted to 1.4 billion tweets containing links to top news, videos, images, and petitions sites
- Aggregated tweets by URL, resulting in 1 billion distinct "events"
- Crawled friend list of each adopter
- Inferred "who got what from whom" to construct diffusion trees
- Characterized size and structure of trees

The Structural Virality of Online Diffusion



Information diffusion

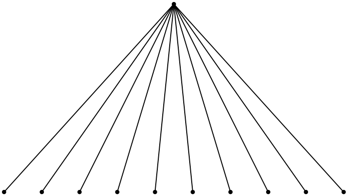
Cascade size distribution

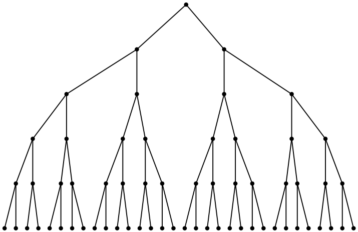


Focus on the rare hits that get at least 100 adoptions

Quantifying structure

Measure the average distance between all pairs of nodes[6]



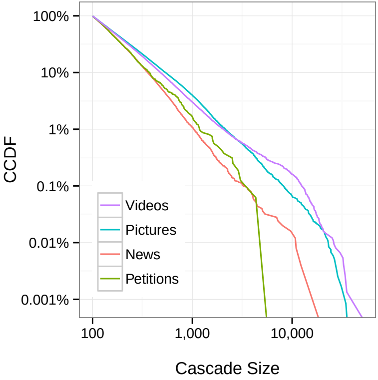


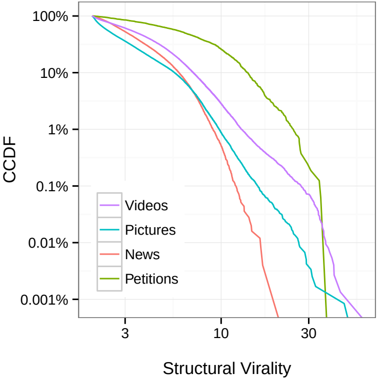
[6]Weiner (1947); correlated with other possible metrics

Information diffusion

Size and virality by category

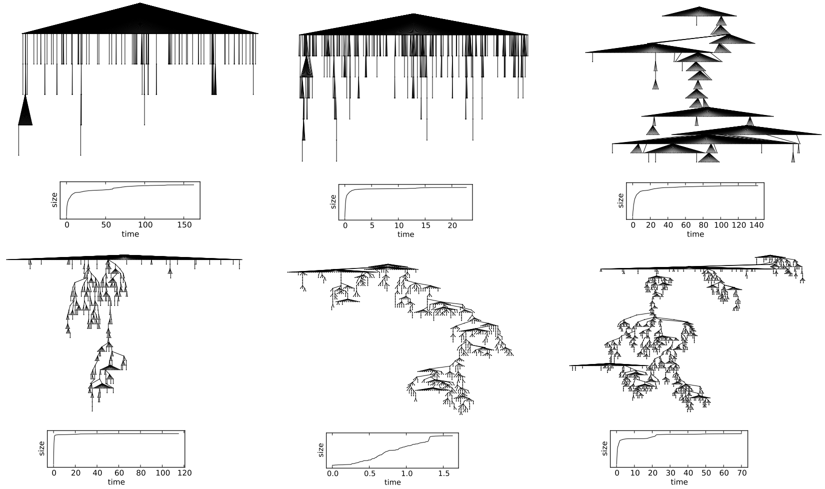
Remarkable structural diversity across across categories





Information diffusion

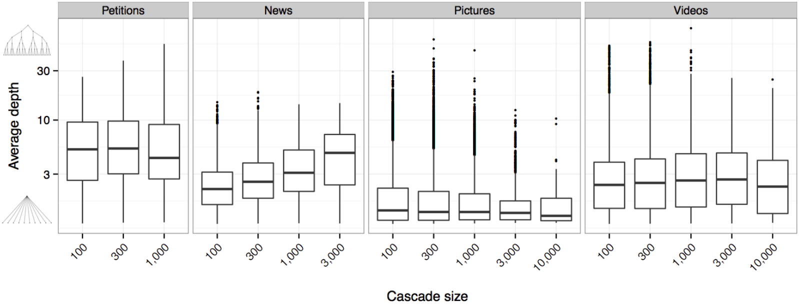
Structural diversity



Information diffusion

Structural diversity

Size is relatively poor predictive of structure



Summary

$$\text{Popular} \neq \text{Viral}$$

Information diffusion

Summary

- Most cascades fail, resulting in fewer than two adoptions, on average
- Of the hits that do succeed, we observe a wide range of diverse diffusion structures
- It's difficult to say how something spread given only its popularity
- "The structural virality of online diffusion", Anderson, Goel, Hofman & Watts (Management Science 2015)

1. Ask good questions

There's nothing interesting in the data without them

2. Think before you code

5 minutes at the whiteboard is worth an hour at the keyboard

3. Keep the answers simple

Exploratory data analysis and linear models go a long way