CSE 258 – Lecture 17

Web Mining and Recommender Systems

More temporal dynamics

This week

Temporal models

This week we'll look back on some of the topics already covered in this class, and see how they can be adapted to make use of **temporal** information

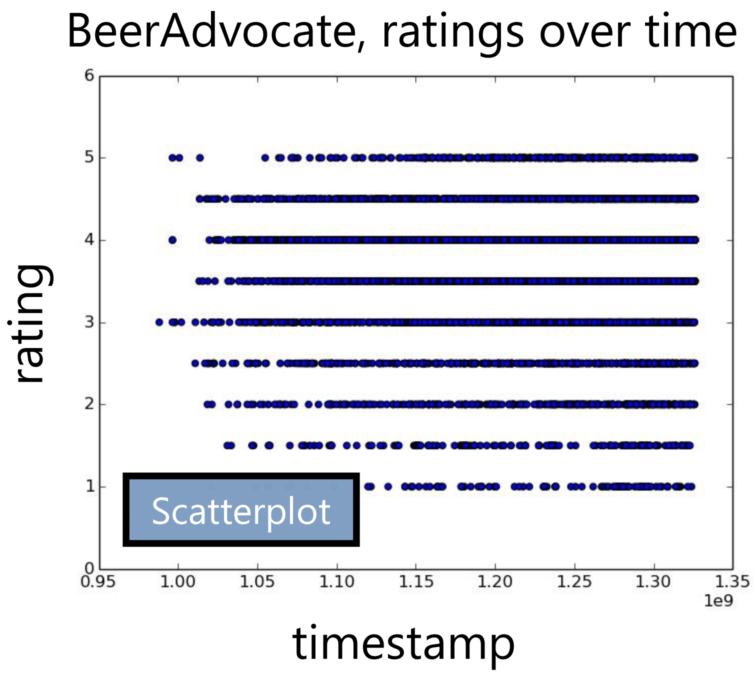
1. **Regression** – sliding windows and autoregression
2. **Classification** – dynamic time-warping
3. **Dimensionality reduction** - ?
4. **Recommender systems** – some results from Koren

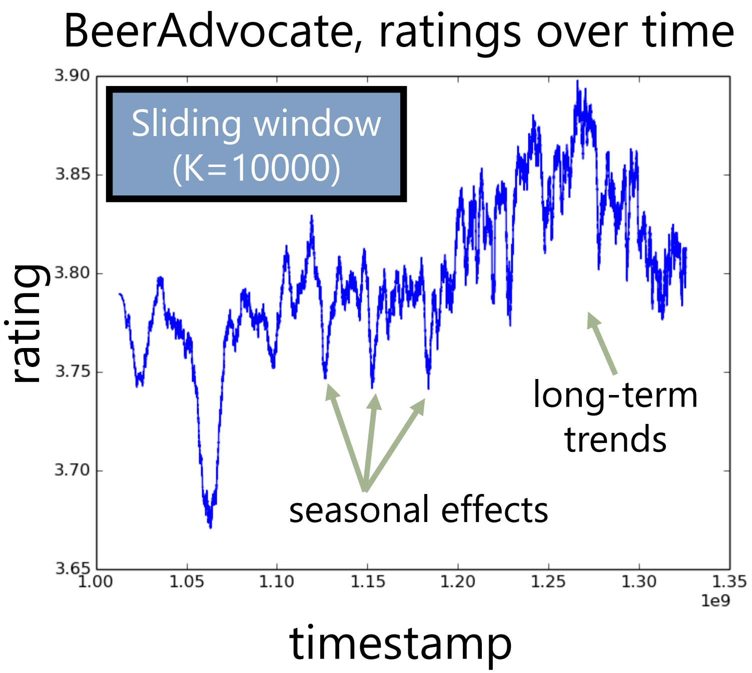
Today:

1. **Text mining** – "Topics over Time"
2. **Social networks** – densification over time

Monday: Time-series regression

Also useful to plot data:





Code on:
http://jmcauley.ucsd.edu/code/week10.py

Monday: Time-series classification

As you recall…
The longest-common subsequence algorithm is a standard dynamic programming problem

	-	A	G	C	A	T
-	0	0	0	0	0	0
G	0	↑← 0	↖ 1	← 1	← 1	← 1
A	0	↖ 1	↑← 1	↑← 1	↖ 2	← 2
C	0	↑ 1	↑← 1	↖ 2	↑← 2	↑← 2

1st sequence

2nd sequence

← = optimal move is to delete from 1st sequence

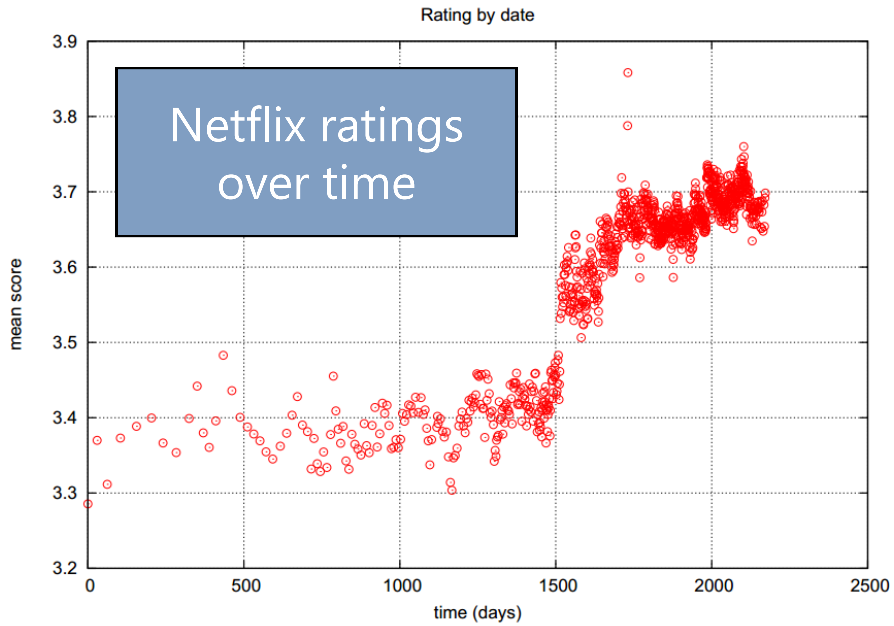
↑ = optimal move is to delete from 2nd sequence

↑← = either deletion is equally optimal

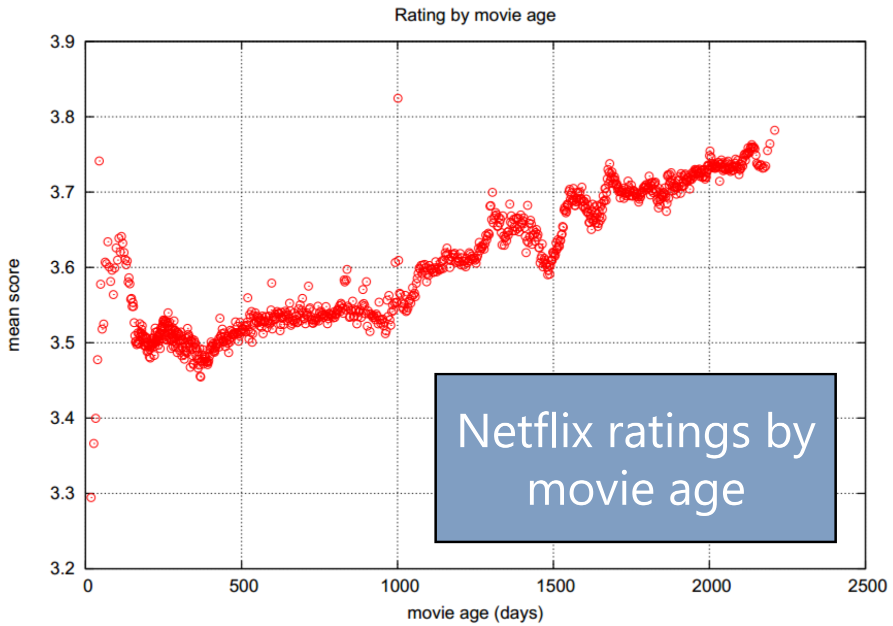
↖ = optimal move is a match

Monday: Temporal recommendation

To build a reliable system (and to win the Netflix prize!) we need to account for **temporal dynamics:**



(Netflix changed their interface)



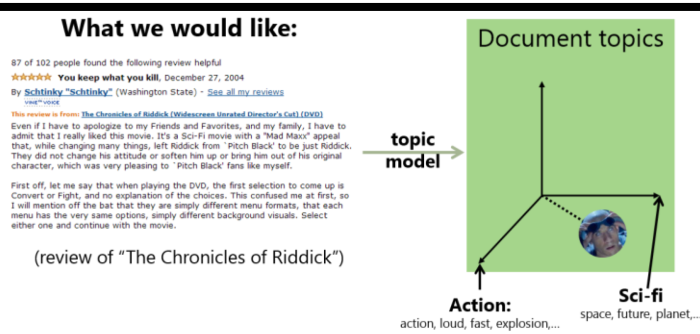
(People tend to give higher ratings to older movies)

Figure from Koren: "Collaborative Filtering with Temporal Dynamics" (KDD 2009)

Week 5/7: Text

yeast and minimal red body thick light a Flavor sugar strong quad. grape over is molasses lace the low and caramel fruit Minimal start and toffee. dark plum, dark brown Actually, alcohol Dark oak, nice vanilla, has brown of a with presence. light carbonation. bready from retention. with finish. with and this and plum and head, fruit, low a Excellent raisin aroma Medium tan

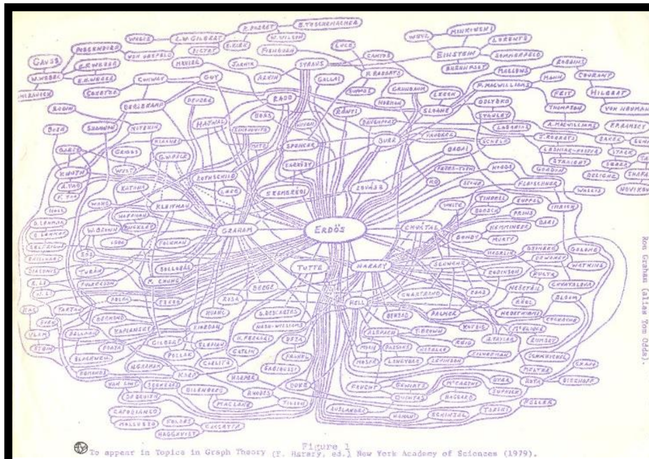
Bags-of-Words



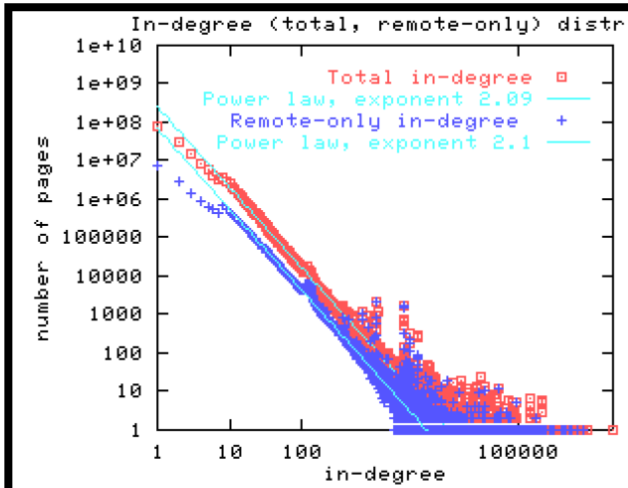
Topic models

Sentiment analysis

8. Social networks

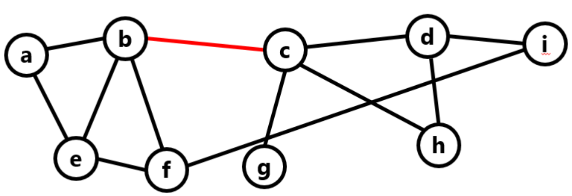


Hubs & authorities



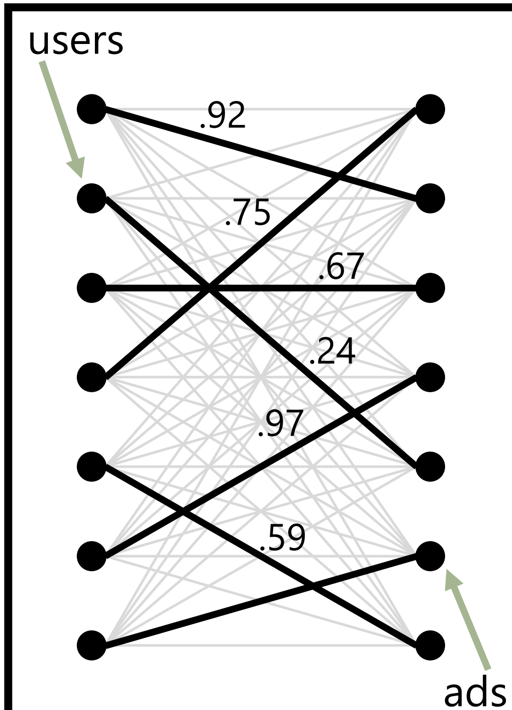
Power laws

Small-world phenomena



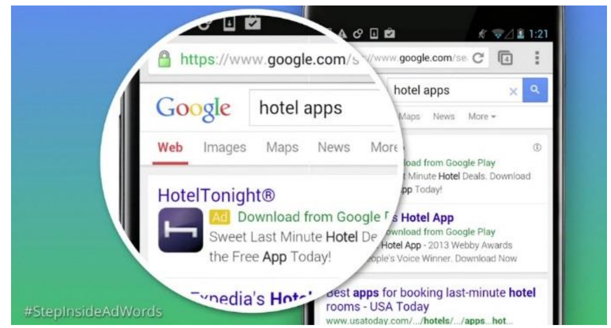
Strong & weak ties

9. Advertising



Matching problems

AdWords



Bandit algorithms

CSE 258 – Lecture 17

Web Mining and Recommender Systems

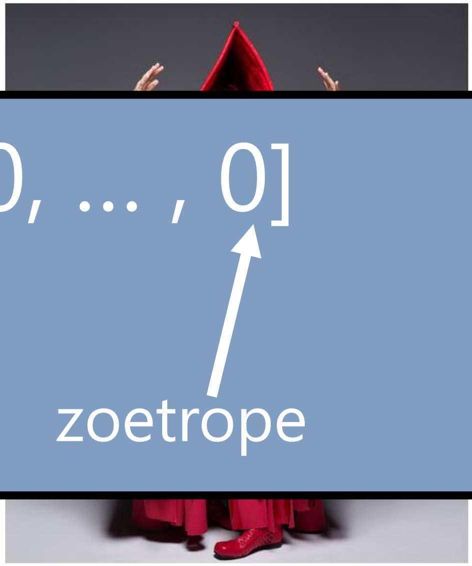
Temporal dynamics of text

Bag-of-Words representations of text:

The Peculiar Genius of Bjork

CULTURE | BY EMILY WITT | JANUARY 23, 2015 11:30 AM

Solo musician or master collaborator? For her new album, Bjork has merged the two sides of her artistry to create a new experience of music — again.



F_text = [150, 0, 0, 0, 0, 0, ... , 0]

a

aardvark

zoetrope

musician, who creates her music in an emotional cocoon, tinkering with technologies, concepts and feelings; and Bjork the producer and curator, who seeks out

Latent Dirichlet Allocation

In week 5/7, we tried to develop low-dimensional representations of documents:

What we would like:

87 of 102 people found the following review helpful

★★★★★ **You keep what you kill**, December 27, 2004

By Schtinky "Schtinky" (Washington State) - See all my reviews

VINE™ VOICE

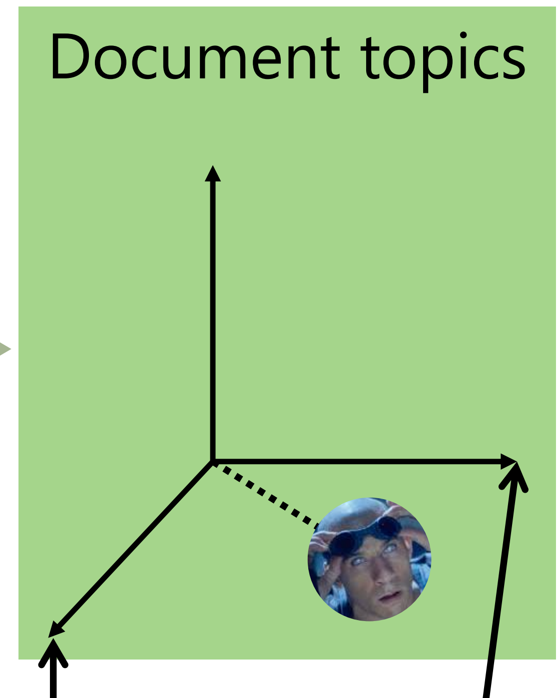
This review is from: The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)

Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from `Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to `Pitch Black' fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of "The Chronicles of Riddick")

topic model

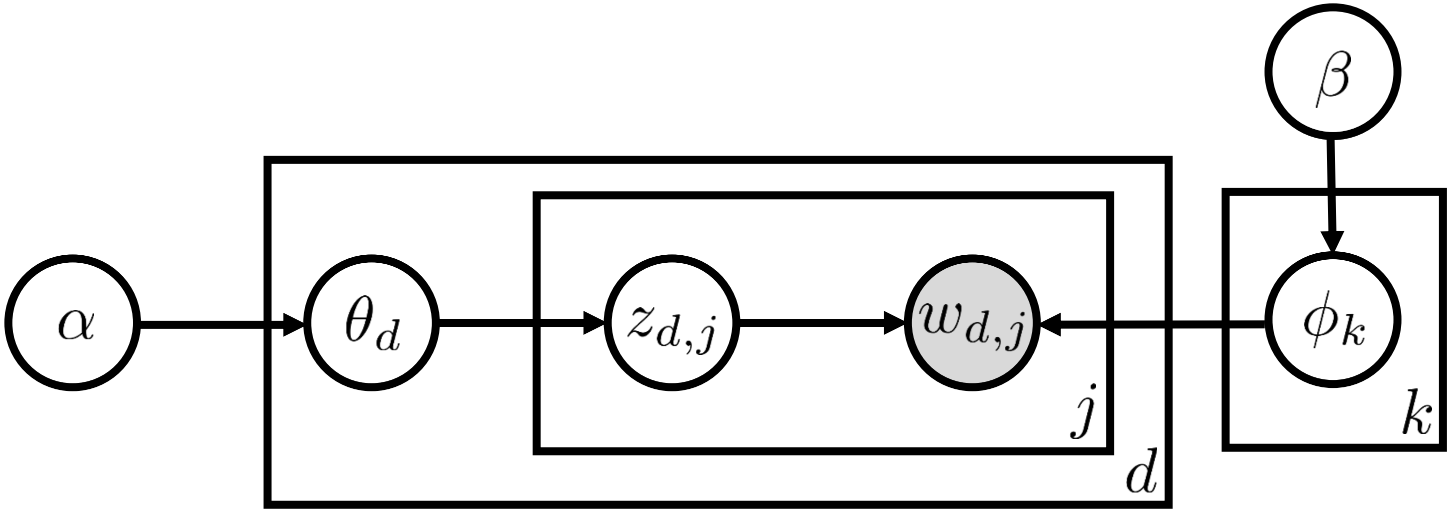


Action:
action, loud, fast, explosion,...

Sci-fi
space, future, planet,...

Latent Dirichlet Allocation

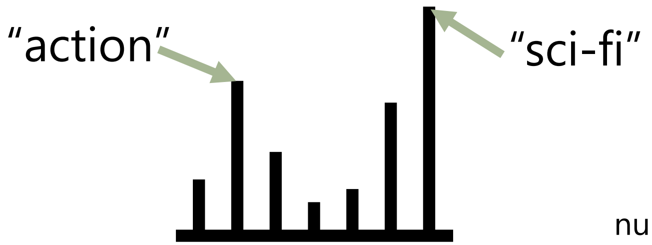
We saw how **LDA** can be used to describe documents in terms of **topics**



- Each document has a **topic vector** (a stochastic vector describing the fraction of words that discuss each topic)
- Each topic has a **word vector** (a stochastic vector describing how often a particular word is used in that topic)

Latent Dirichlet Allocation

Topics and documents are **both** described using stochastic vectors:

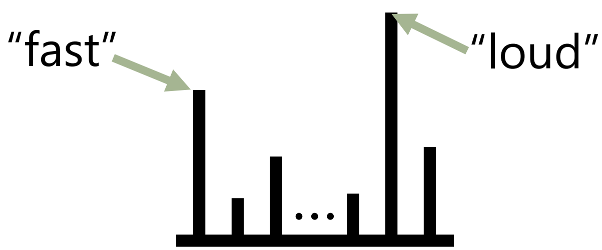


$\theta_{\text{pitch black}}$

Each document has a **topic distribution** which is a mixture over the topics it discusses

number of topics

$\theta_d \in \Delta^K$ i.e., $\forall_d \sum_k \theta_{d,k} = 1$



ϕ_{action}

Each topic has a **word distribution** which is a mixture over the words it discusses

number of words

$\phi_k \in \Delta^D$ i.e., $\forall_k \sum_w \phi_{k,w} = 1$

Latent Dirichlet Allocation

Topics over Time (Wang & McCallum, 2006) is an approach to incorporate temporal information into topic models

e.g.

- The topics discussed in conference proceedings progressed from neural networks, towards SVMs and structured prediction (and back to neural networks)
- The topics used in political discourse now cover science and technology more than they did in the 1700s
- With in an institution, e-mails will discuss different topics (e.g. recruiting, conference deadlines) at different times of the year

Latent Dirichlet Allocation

Topics over Time (Wang & McCallum, 2006) is an approach to incorporate temporal information into topic models

The ToT model is similar to LDA with one addition:

1. For each topic K, draw a word vector \phi_k from Dir.(\beta)
2. For each document d, draw a topic vector \theta_d from Dir.(\alpha)
3. For each word position i:
    1. draw a topic z_{di} from multinomial \theta_d
    2. draw a word w_{di} from multinomial \phi_{z_{di}}
    3. **draw a timestamp t_{di} from Beta(\psi_{z_{di}})**

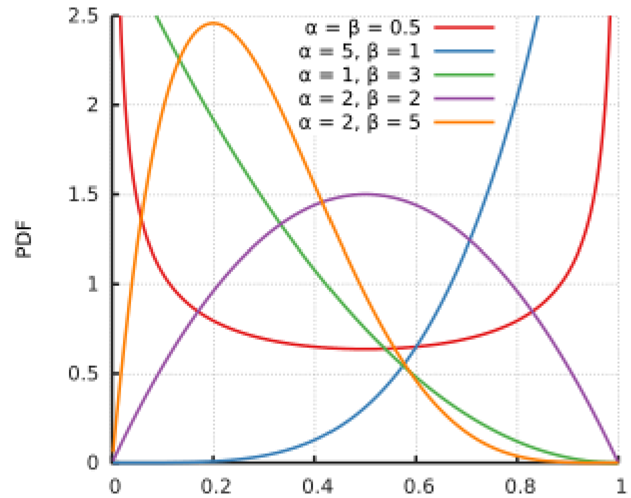
Latent Dirichlet Allocation

Topics over Time (Wang & McCallum, 2006) is an approach to incorporate temporal information into topic models

3.3. draw a timestamp t_{di} from Beta(\psi_{z_{di}})

- There is now one Beta distribution **per topic**
- Inference is still done by Gibbs sampling, with an outer loop to update the Beta distribution parameters

Beta distributions are a flexible family of distributions that can capture several types of behavior – e.g. gradual increase, gradual decline, or temporary “bursts”



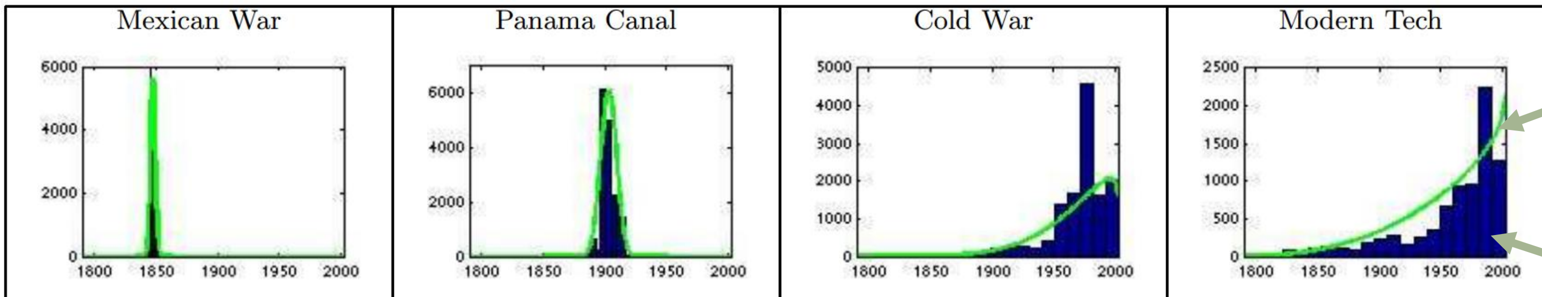
p.d.f.:

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

Latent Dirichlet Allocation

Results:

Political addresses – the model seems to capture realistic “bursty” and gradually emerging topics



Mexican War		Panama Canal		Cold War		Modern Tech	
states	0.02032	government	0.02928	world	0.01875	energy	0.03902
mexico	0.01832	united	0.02132	states	0.01717	national	0.01534
government	0.01670	states	0.02067	security	0.01710	development	0.01448
united	0.01521	islands	0.01167	soviet	0.01664	space	0.01436
war	0.01059	canal	0.01014	united	0.01491	science	0.01227
congress	0.00951	american	0.00872	nuclear	0.01454	technology	0.01227
country	0.00906	cuba	0.00834	peace	0.01408	oil	0.01178
texas	0.00852	made	0.00747	nations	0.01069	make	0.00994
made	0.00727	general	0.00731	international	0.01024	effort	0.00969
great	0.00611	war	0.00660	america	0.00987	administration	0.00957

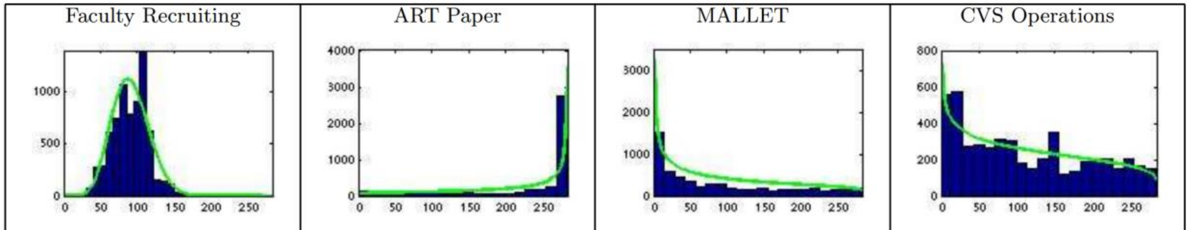
fitted Beta distrbution

assignments to this topic

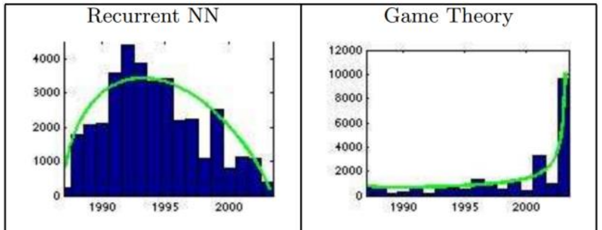
Latent Dirichlet Allocation

Results:

e-mails & conference proceedings



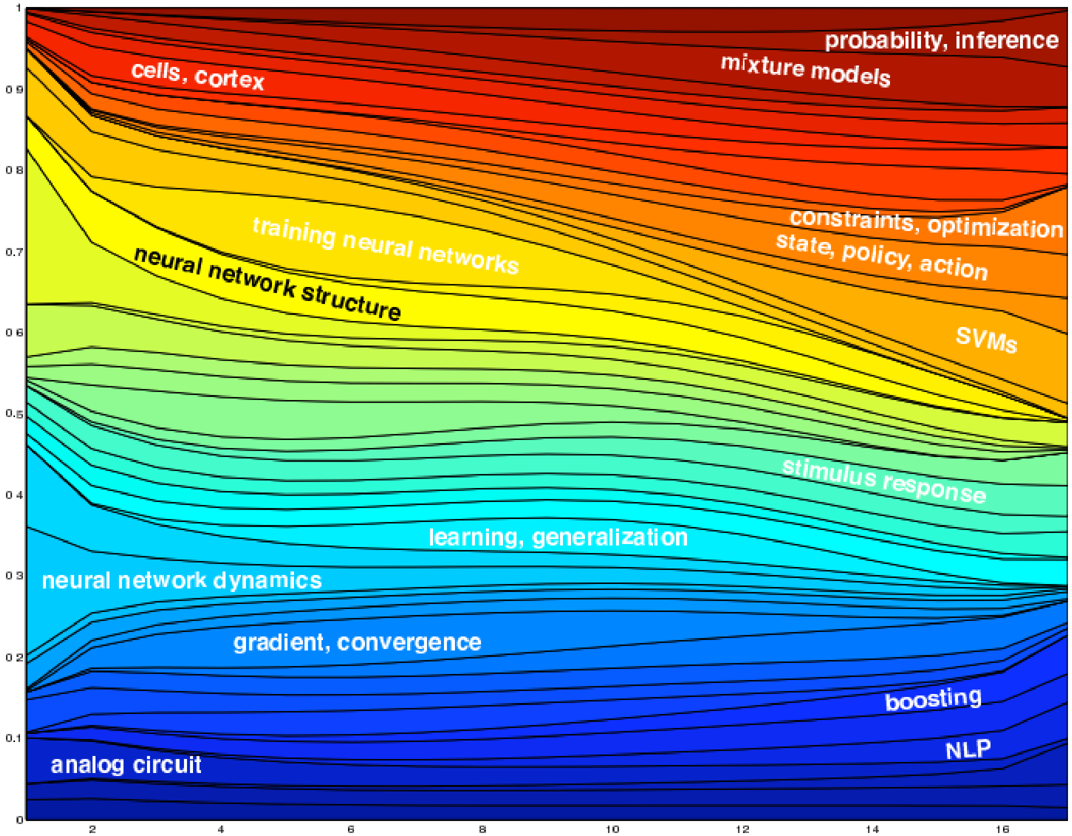
Faculty Recruiting		ART Paper		MALLET		CVS Operations	
cs	0.03572	xuerui	0.02113	code	0.05668	check	0.04473
april	0.02724	data	0.01814	files	0.04212	page	0.04070
faculty	0.02341	word	0.01601	mallet	0.04073	version	0.03828
david	0.02012	research	0.01408	java	0.03085	cvs	0.03587
lunch	0.01766	topic	0.01366	file	0.02947	add	0.03083
schedule	0.01656	model	0.01238	al	0.02479	update	0.02539
candidate	0.01560	andres	0.01238	directory	0.02080	latest	0.02519
talk	0.01355	sample	0.01152	version	0.01664	updated	0.02317
bruce	0.01273	enron	0.01067	pdf	0.01421	checked	0.02277
visit	0.01232	dataset	0.00960	bug	0.01352	change	0.02156



Recurrent NN		Game Theory	
state	0.05963	game	0.02850
recurrent	0.03765	strategy	0.02378
sequence	0.03616	play	0.01490
sequences	0.02462	games	0.01473
time	0.02402	player	0.01451
states	0.02057	agents	0.01346
transition	0.01300	expert	0.01281
finite	0.01242	strategies	0.01123
length	0.01154	opponent...	0.01088
strings	0.01013	nash	0.00848

Latent Dirichlet Allocation

Results:
conference proceedings (NIPS)



Relative weights of various topics in 17 years of NIPS proceedings

Questions?

Further reading:
"Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends"
(Wang & McCallum, 2006)
http://people.cs.umass.edu/~mccallum/papers/tot-kdd06.pdf

CSE 258 – Lecture 17

Web Mining and Recommender Systems

Temporal dynamics of social networks

Week 8

How can we **characterize, model,** and **reason about** the structure of social networks?

1. Models of network structure
2. Power-laws and scale-free networks, “rich-get-richer” phenomena
3. Triadic closure and “the strength of weak ties”
4. Small-world phenomena
5. Hubs & Authorities; PageRank

Temporal dynamics of social networks

Two weeks ago we saw some processes that model the generation of social and information networks

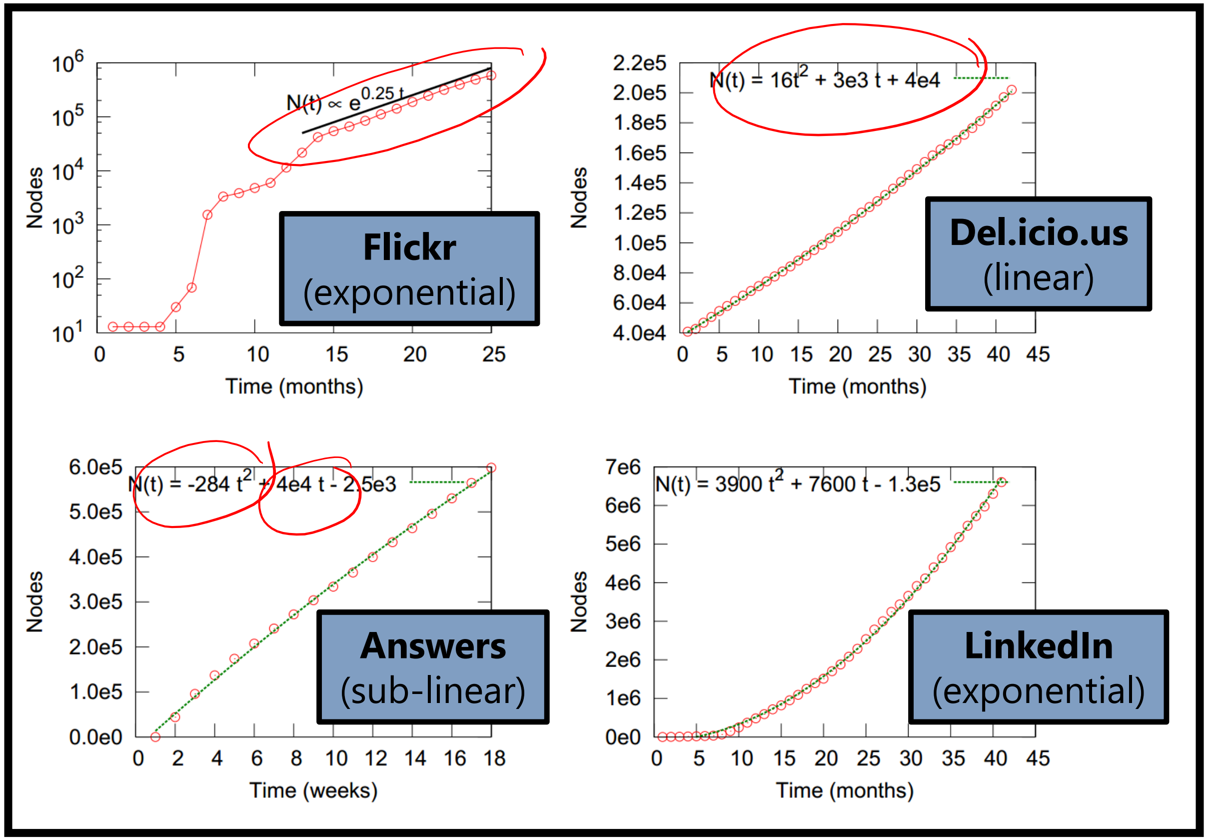
- Power-laws & small worlds
- Random graph models

These were all defined with a “static” network in mind. But if we observe the **order** in which edges were created, we can study how these phenomena change as a function of time

First, let’s look at “microscopic” evolution, i.e., evolution in terms of individual nodes in the network

Temporal dynamics of social networks

Q1: How do networks grow in terms of the number of nodes over time?



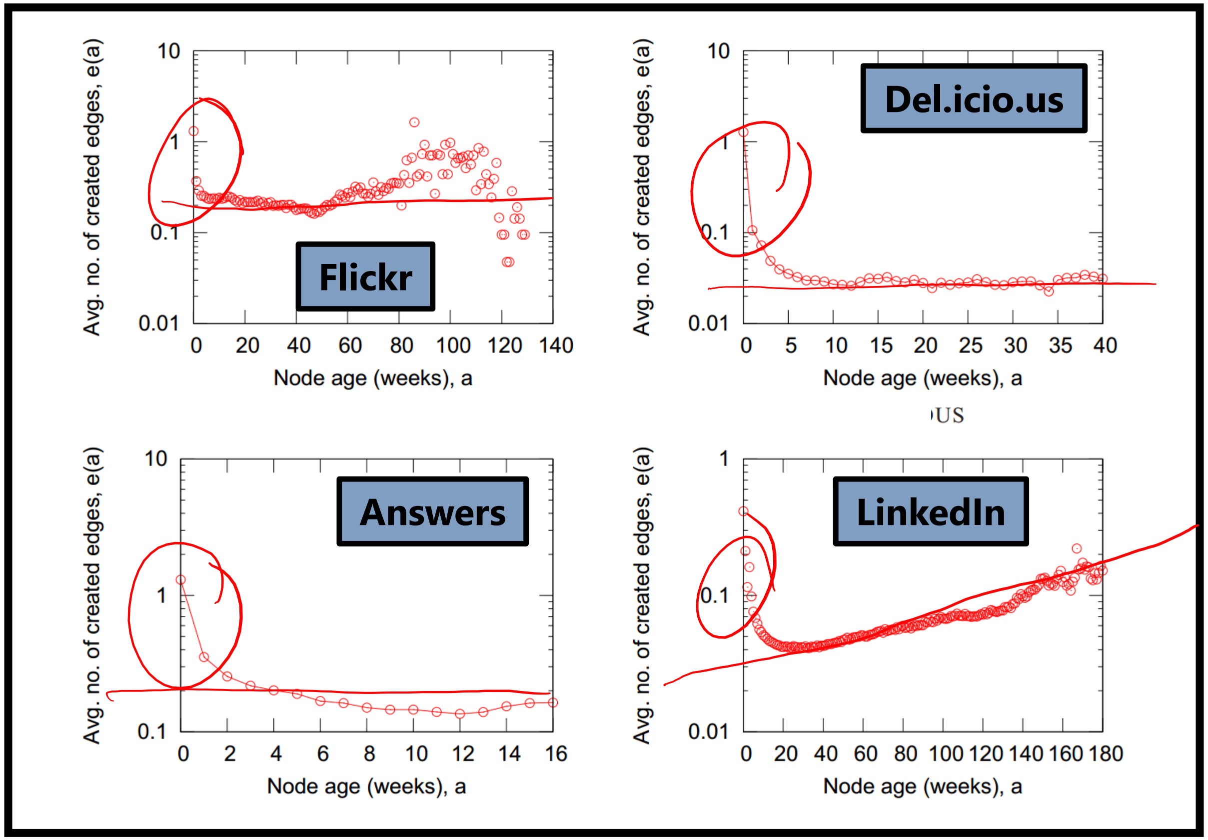
(from Leskovec, 2008 (CMU Thesis))

A: Doesn't seem to be an obvious trend, so what **do** networks have in common as they evolve?

Temporal dynamics of social networks

Q2: When do nodes create links?

- x-axis is the age of the nodes
- y-axis is the number of edges created at that age

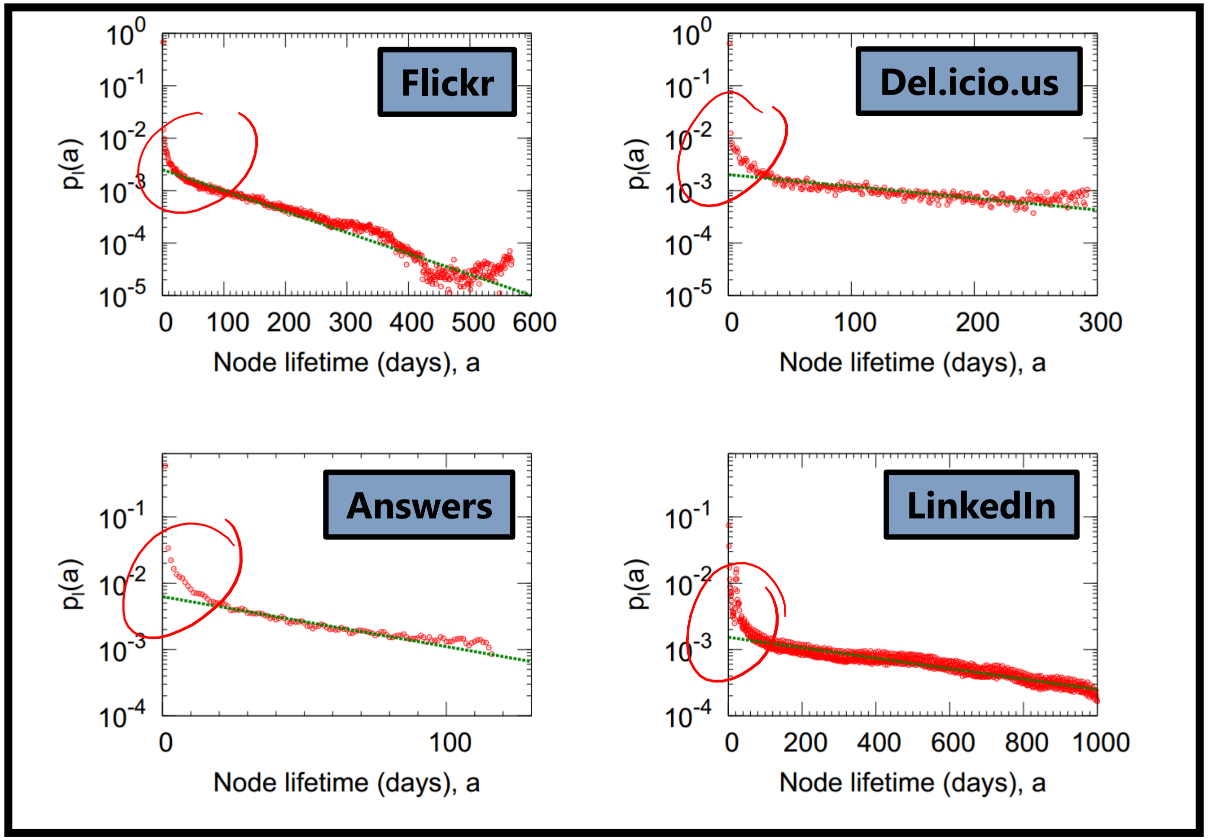


A: In most networks there's a "burst" of initial edge creation which gradually flattens out.
Very different behavior on LinkedIn (guesses as to why?)

Temporal dynamics of social networks

Q3: How long do nodes "live"?

- x-axis is the diff. between date of last and first edge creation
- y-axis is the frequency



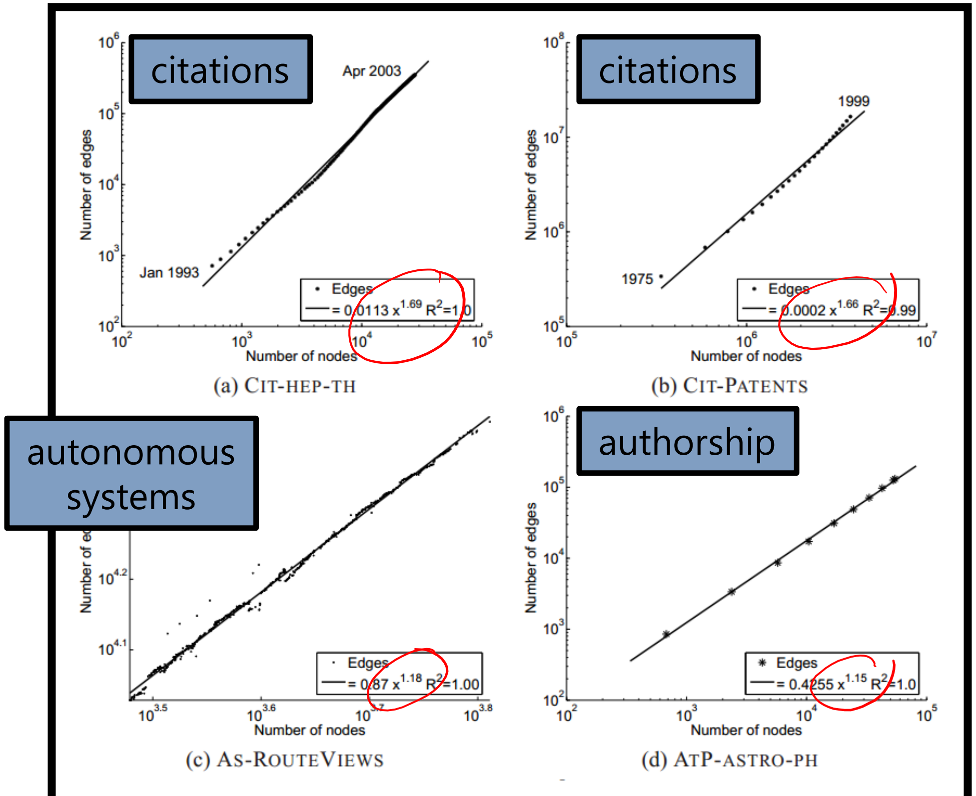
A: Node lifetimes follow a power-law: many many nodes are shortlived, with a long-tail of older nodes

Temporal dynamics of social networks

What about "macroscopic" evolution, i.e., how do global properties of networks change over time?

Q1: How does the # of nodes relate to the # of edges?

$E = cN^{\alpha}$



- A few more networks: citations, authorship, and autonomous systems (and some others, not shown)
- **A:** Seems to be linear (on a log-log plot) **but** the number of edges grows **faster** than the number of nodes as a function of time

Temporal dynamics of social networks

Q1: How does the # of nodes relate to the # of edges?
A: seems to behave like

$$E(t) \propto N(t)^a$$

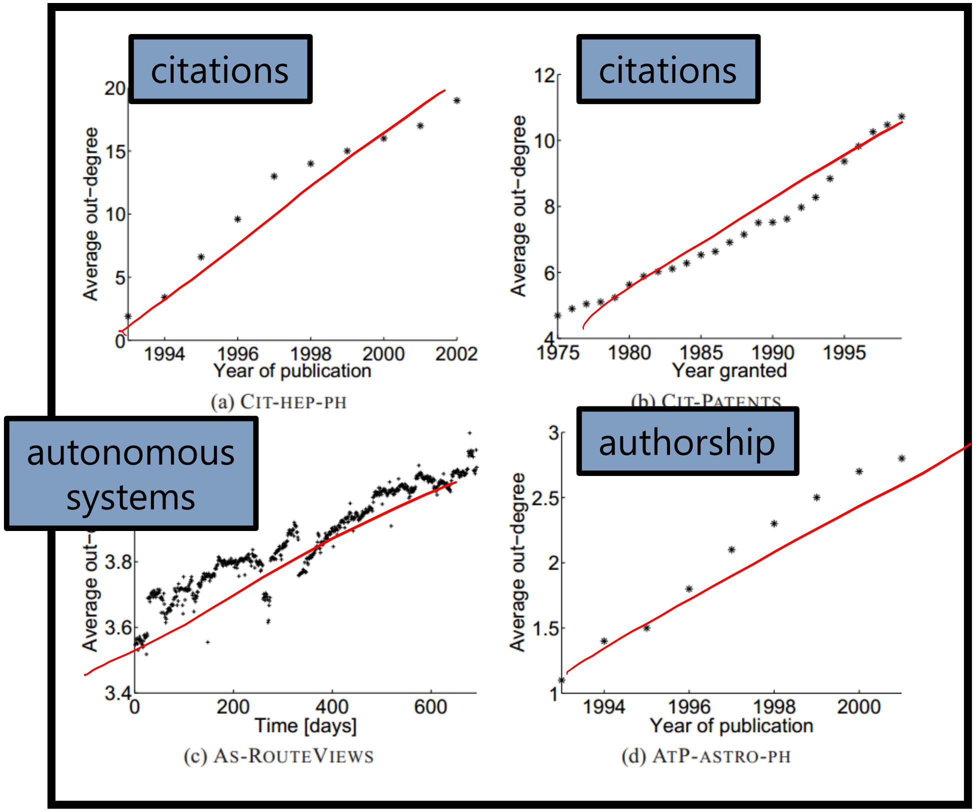
where

$$1 \leq a \leq 2$$

- a = 1 would correspond to **constant** out-degree – which is what we might traditionally assume
- a = 2 would correspond to the graph being fully connected
- What seems to be the case from the previous examples is that a > 1 – the number of edges grows faster than the number of nodes

Temporal dynamics of social networks

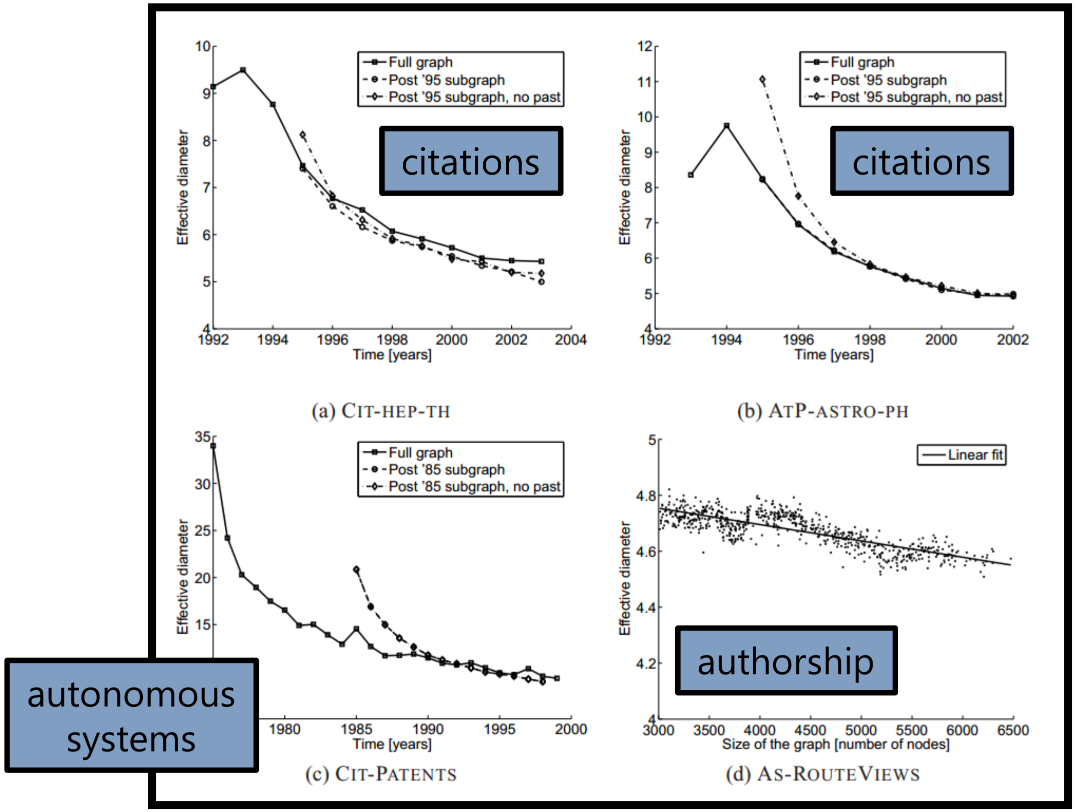
Q2: How does the degree change over time?



- **A:** The average out-degree **increases** over time

Temporal dynamics of social networks

Q3: If the network becomes **denser**, what happens to the (effective) diameter?



- **A:** The diameter seems to decrease
- In other words, the network becomes **more** of a small world as the number of nodes increases

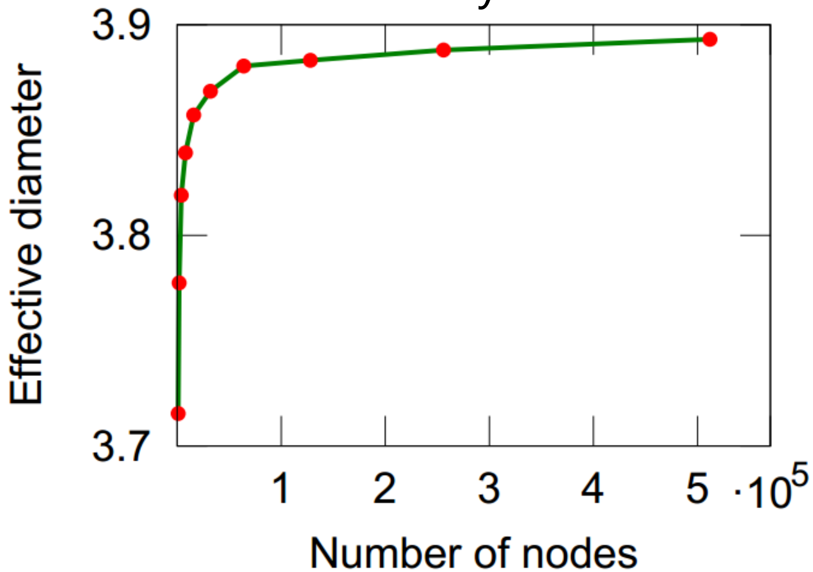
Temporal dynamics of social networks

Q4: Is this something that **must** happen – i.e., if the number of edges increases faster than the number of nodes, does that mean that the diameter must decrease?

A: Let's construct random graphs (with a > 1) to test this:

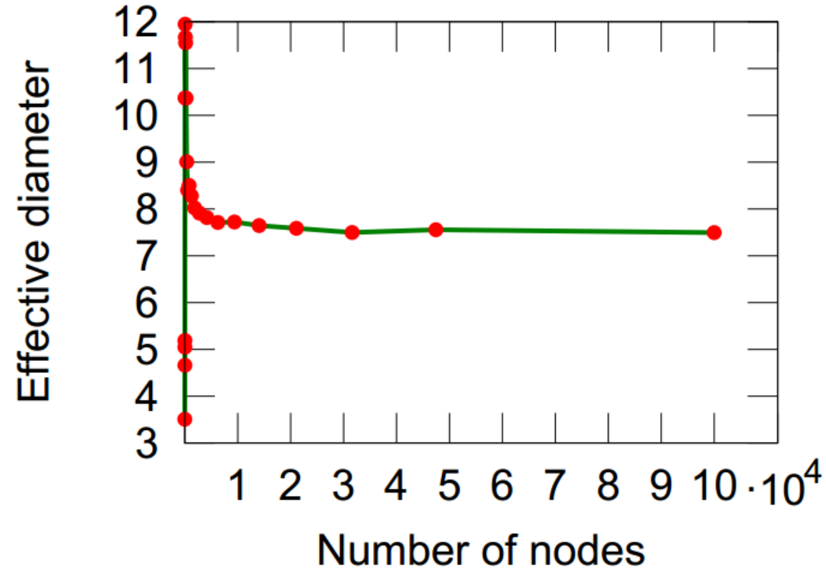
$E = N^{1.3}$

Erdos-Renyi – a = 1.3



$E = N^{1.2}$

Pref. attachment model – a = 1.2

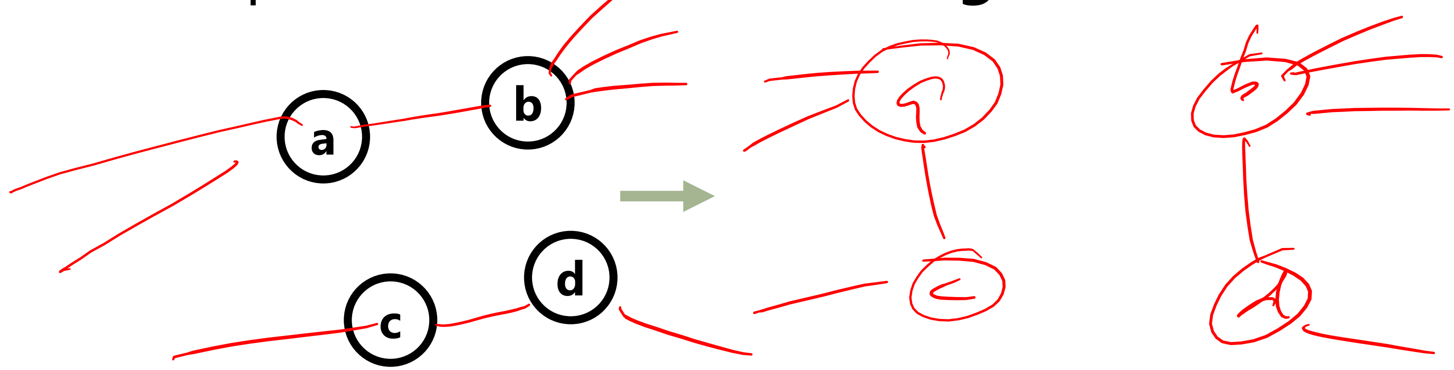


Temporal dynamics of social networks

So, a decreasing diameter is **not** a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model

Q5: is the degree distribution of the nodes sufficient to explain the observed phenomenon?

A: Let’s perform **random rewiring** to test this

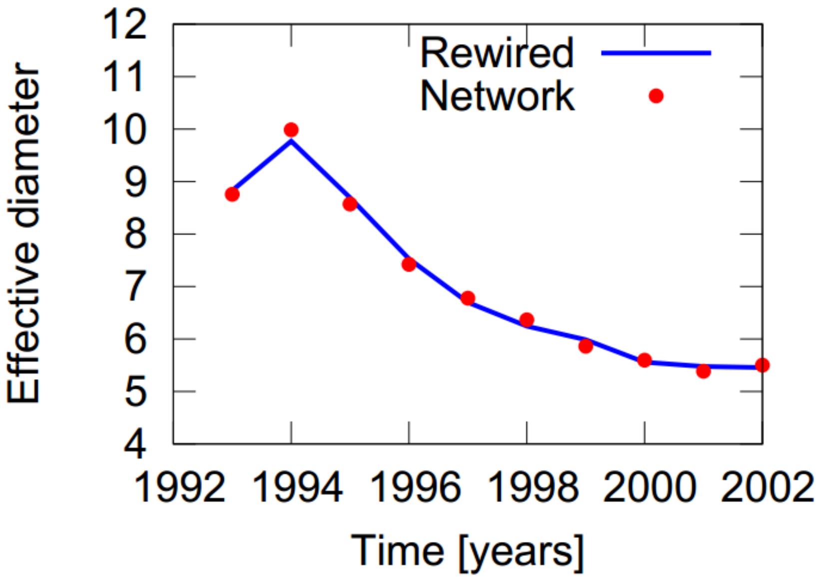


random rewiring preserves the degree distribution, and randomly samples amongst networks with observed degree distribution

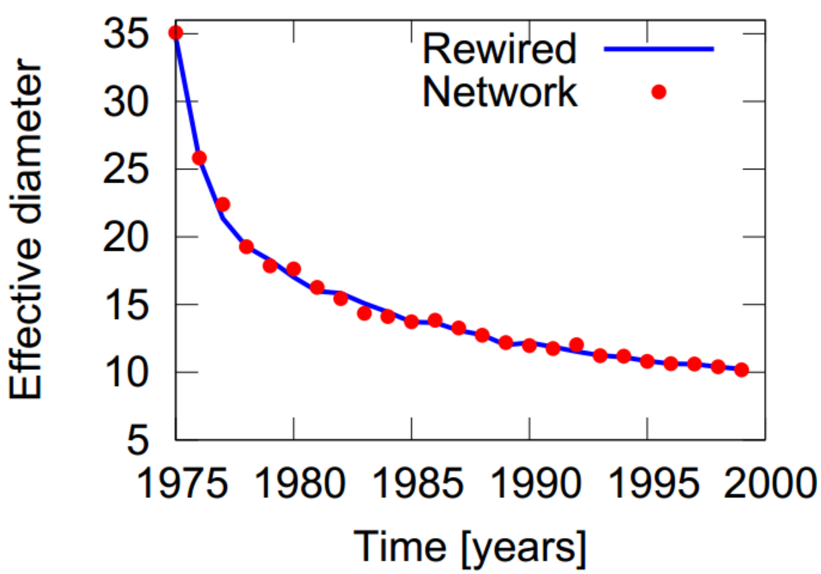
Temporal dynamics of social networks

So, a decreasing diameter is **not** a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model

Q5: is the degree distribution of the nodes sufficient to explain the observed phenomenon?



(c) Affiliation network (ATP-ASTRO-PH)



(d) US patent citation network (CIT-PATENTS)

Temporal dynamics of social networks

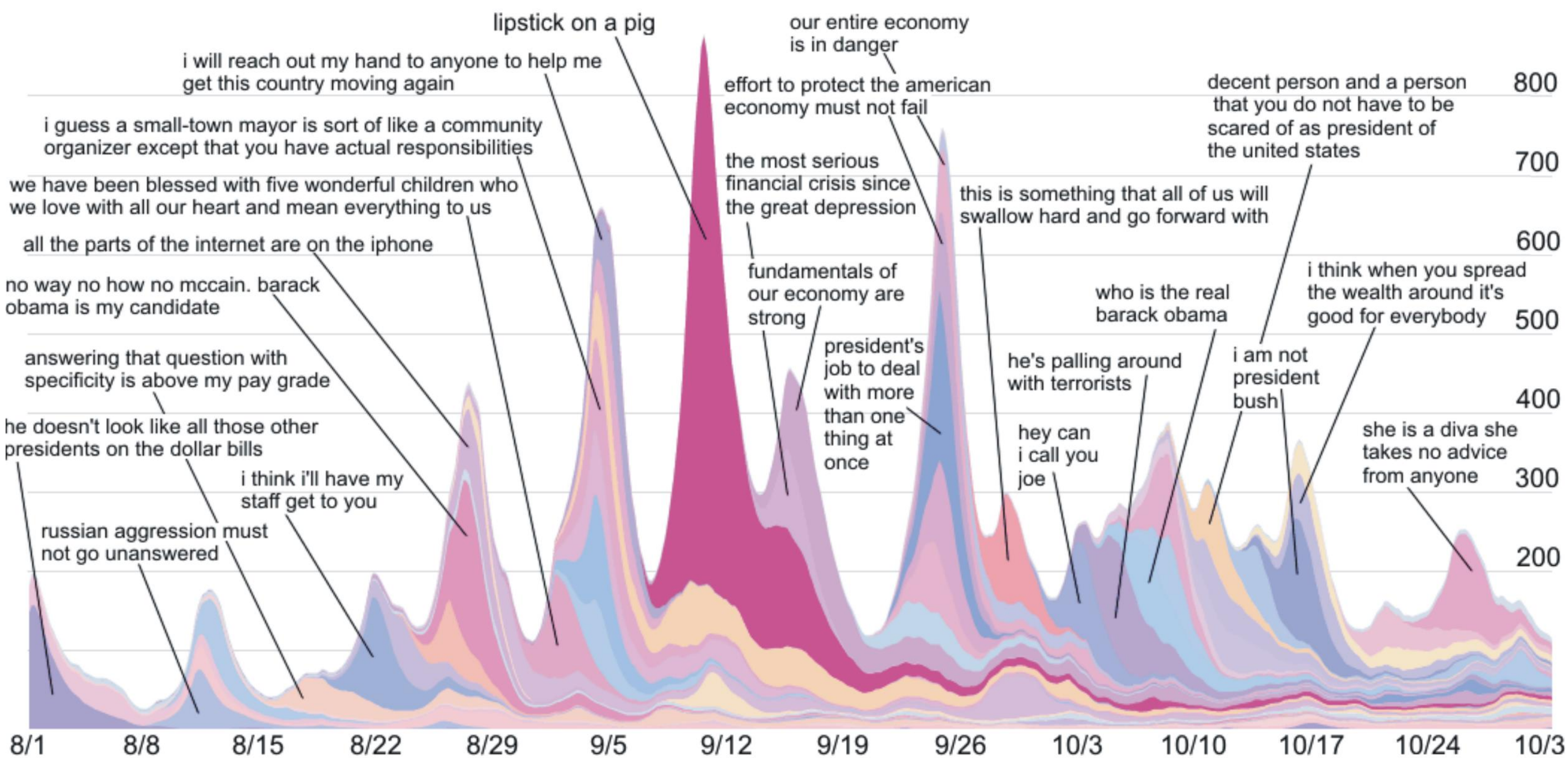
So, a decreasing diameter is **not** a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model

Q5: is the degree distribution of the nodes sufficient to explain the observed phenomenon?

A: Yes! The fact that real-world networks seem to have decreasing diameter over time can be explained as a result of their degree distribution **and** the fact that the number of edges grows faster than the number of nodes

Temporal dynamics of social networks

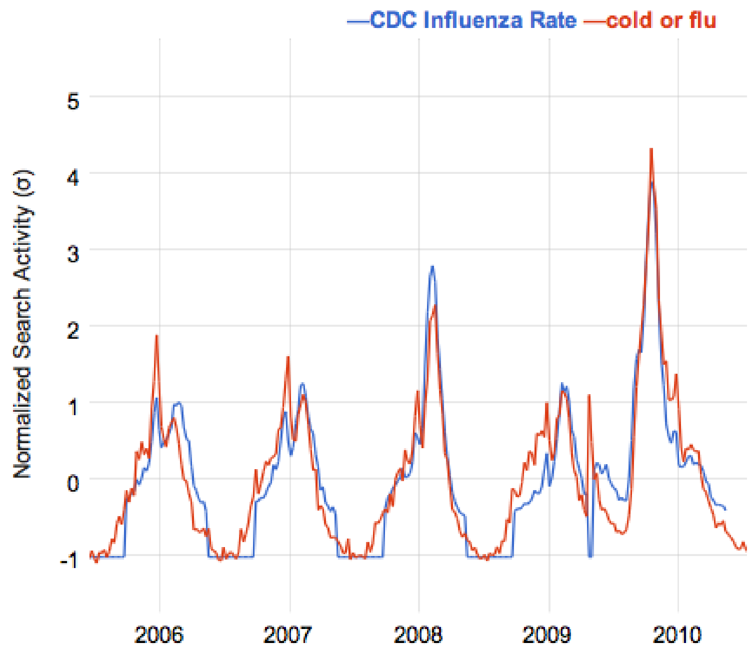
Other interesting topics…



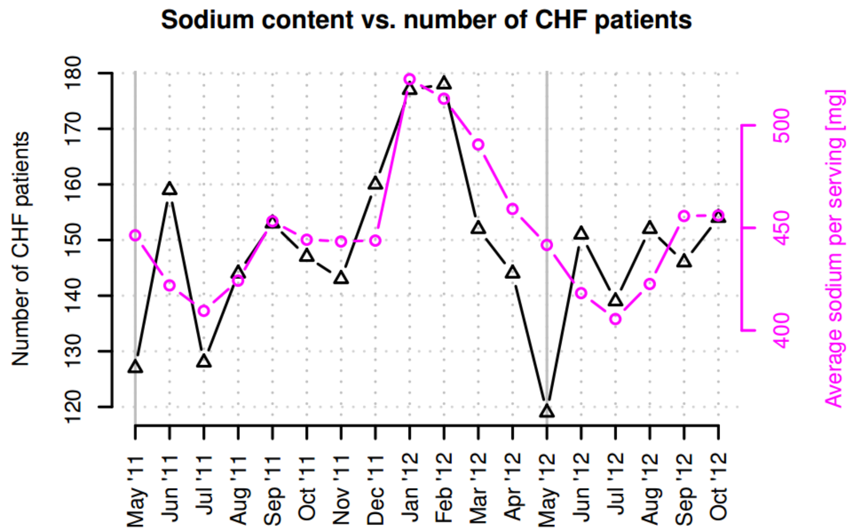
“memetracker”

Temporal dynamics of social networks

Other interesting topics…



Aligning query data with disease data – Google flu trends: https://www.google.org/flutrends/us/#US



Sodium content in recipe searches vs. # of heart failure patients – "From Cookies to Cooks" (West et al. 2013): http://infolab.stanford.edu/~west1/pubs/West-White-Horvitz_WWW-13.pdf

Questions?

Further reading:

"Dynamics of Large Networks" (most plots from here)
Jure Leskovec, 2008
http://cs.stanford.edu/people/jure/pubs/thesis/jure-thesis.pdf

"Microscopic Evolution of Social Networks"
Leskovec et al. 2008
http://cs.stanford.edu/people/jure/pubs/microEvol-kdd08.pdf

"Graph Evolution: Densification and Shrinking Diameters"
Leskovec et al. 2007
http://cs.stanford.edu/people/jure/pubs/powergrowth-tkdd.pdf

CSE 258 – Lecture 17

Web Mining and Recommender Systems

Some incredible assignments

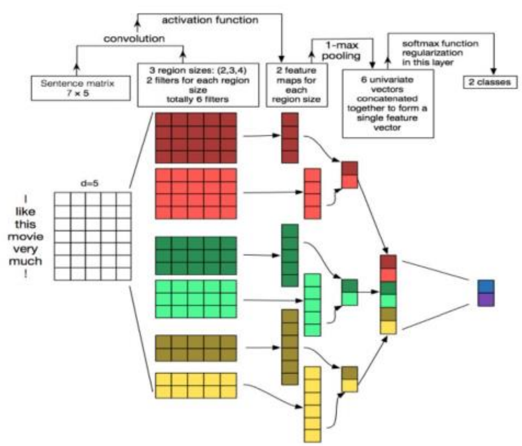
Fake news detection

Grab real and fake news from Kaggle (fake news detection dataset) and *Freedom to Tinker* (real headlines):

abcnews.go.com	55400
bbc.co.uk	37250
breitbart.com	148836
buzzfeed.com	110848
cbsnews.com	87849
chicagotribune.com	33304
chron.com	142965
cnbc.com	30995
cnn.com	74237
forbes.com	20077
foxnews.com	104173
hollywoodreporter.com	36217
huffingtonpost.com	72268
latimes.com	137928
money.cnn.com	171684
nbcnews.com	57621
nypost.com	171295
politico.com	18462
reuters.com	64474
theguardian.com	68642
time.com	199723
usatoday.com	22632
usnews.com	118309
wsj.com	63191

bignuggetnews.com	1
bipartisanreport.com	9
blackagendareport.com	94
blacklistednews.com	100
breitbart.com	100
christiantimesnewspaper.com	1
chronicle.su	5
	...
unz.com	100
usanewsflash.com	6
usanewsinsider.com	3
usapoliticsnow.com	9
usasupreme.com	3
usatwentyfour.com	4
usuncut.com	22
usviewer.com	2
vdare.com	100
veteransnewsnow.com	100
veteranstoday.com	100
vigilantcitizen.com	2
viralliberty.com	3
voltairenet.org	100
vonpress.com	1
wakingtimes.com	41
washingtonsblog.com	100
waterfordwhispersnews.com	100
wearechange.org	100
westernjournalism.com	100
whatreallyhappened.com	14
whydontyoutrythis.com	67
wikileaks.org	8
winningdemocrats.com	2
wnd.com	100
worldnewspolitics.com	1
worldtruth.tv	100
wundergroundmusic.com	2
yournewswire.com	100
zerohedge.com	100

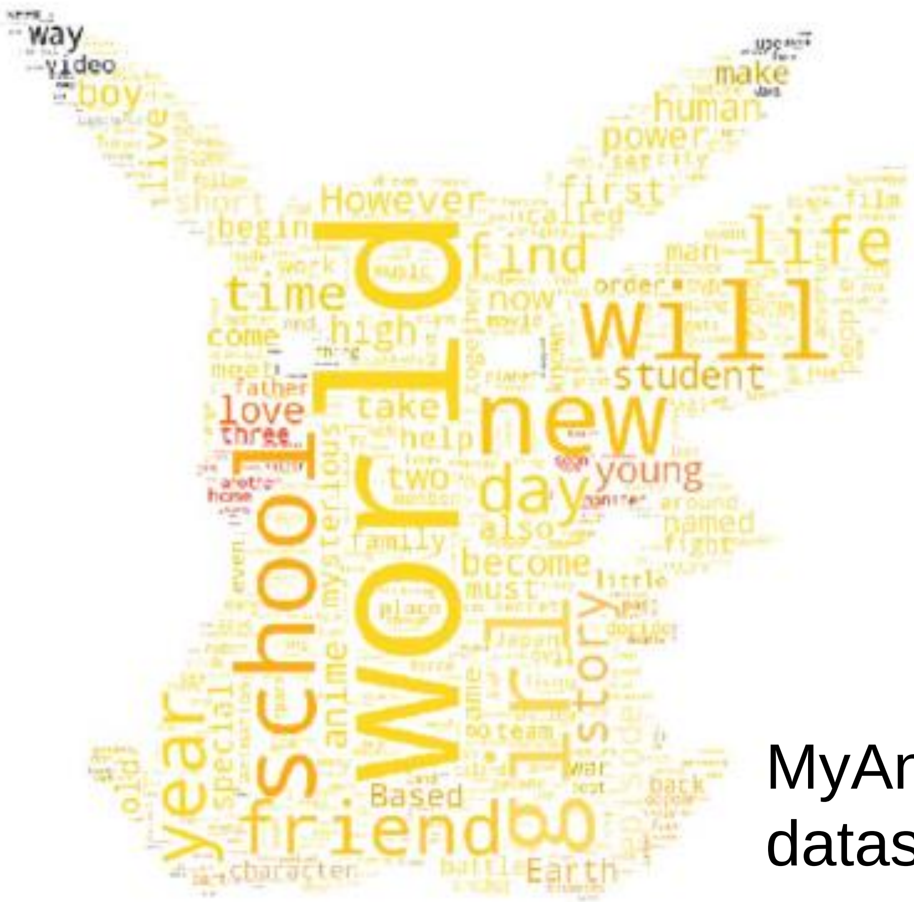
Words from real vs. fake headlines



Extract words and train using a CNN

Jimmy Gia Quach, Shih-Cheng Huang

Anime Recommendation



$$r = \frac{\sum_{v \in V} similarity(u, v) * rating(v, a)}{count(a)}$$

Features	MSE
Always predict average	1.03293792426
Synopsis bag-of-words	0.806018062926
Genre, members, title	0.681102399363
All of the above	0.62533064608

MyAnimeList
dataset from Kaggle

Richard Lin, Daniel Lee

Fine Foods reviews

(a) One-star rating (b) Three-star rating (c) Five-star rating

(a) One-star rating (b) Three-star rating (c) Five-star rating

(a) adjective (b) noun (c) verb

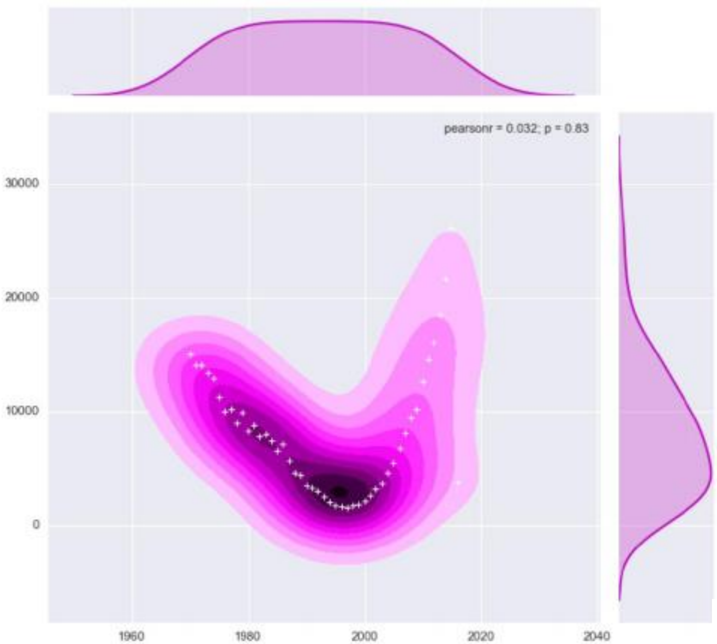
(a) One-star rating (b) Three-star rating (c) Five-star rating

Zhongjian Zhu, Jinhan Zhang, Siqi Qin

Beer reviews

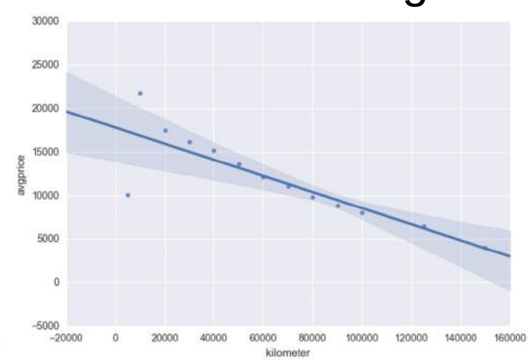
Yunsheng Li, Mengzhi Li, Chenxi Cao

Used car price prediction



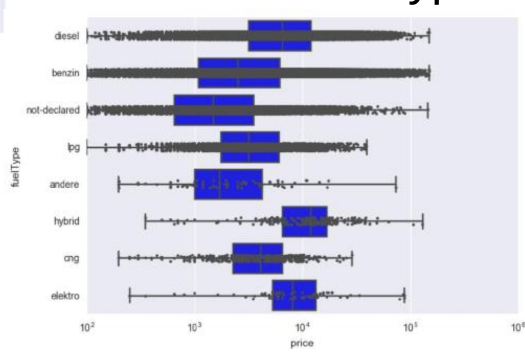
Price vs. registration year

Price vs. mileage



- Type (sedan, van, etc.)
- Mileage
- Age
- PowerPS
- Damage
- Gearbox
- Fuel type

Price vs. fuel type



Features	Train Set Accuracy	Test Set Accuracy
E	0.62770924	0.628140622
D	0.660893404	0.661312692
B	0.685244315	0.686518303
B+E	0.689159007	0.690291888
B+E+D	0.836074827	0.802370585
B+E+D+A	0.88159571	0.830882116
B+E+D+F	0.978870343	0.775907112
B+E+D+C	0.846331237	0.803096275

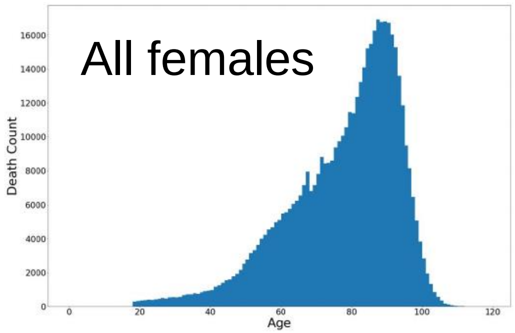
Kaggle used cars dataset (370,000 instances)

Xinyuan Zhang, Changtong Qiu, Zhiye Zhang

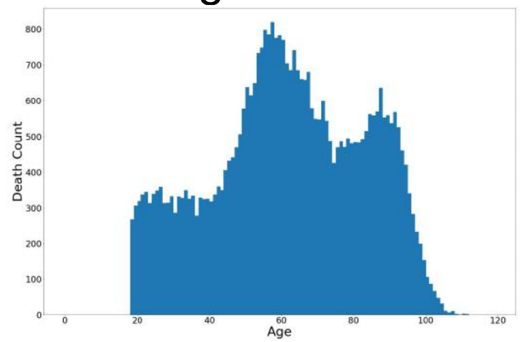
Death clock

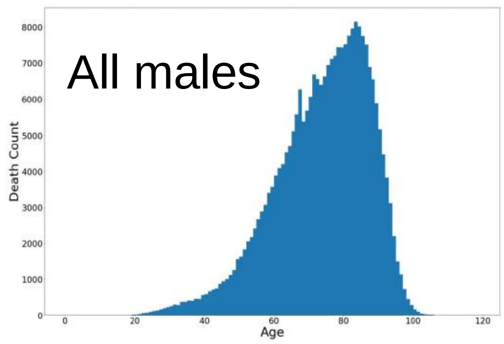
CDC Mortality Dataset (2.1 million instances)

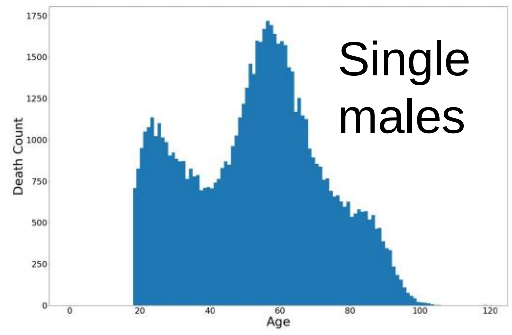
Models	Feature List	MSE
Best	Linear Regression using: {All above features} and 2,100,000 deaths	142.43
2nd	Linear Regression using: {All above features except *Activity Code* features} and 2,100,000 deaths	142.89
3rd	Linear Regression using: {All above features except *Resident Code* features} and 2,100,000 deaths	143.52
4th	Linear Regression using: {All above features except *Education Level* features} and 2,100,000 deaths	144.24
5th	Linear Regression using: {All above features except *Marriage* features} and 2,100,000 deaths	186.49
6th	SVM using {All above features} and 50,000 deaths	178.20
Baseline	Mean age at death	270.25

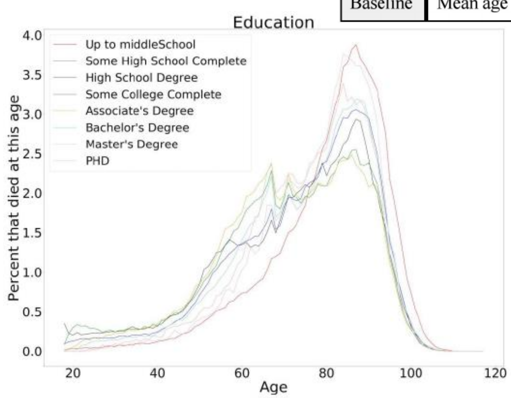


Single females





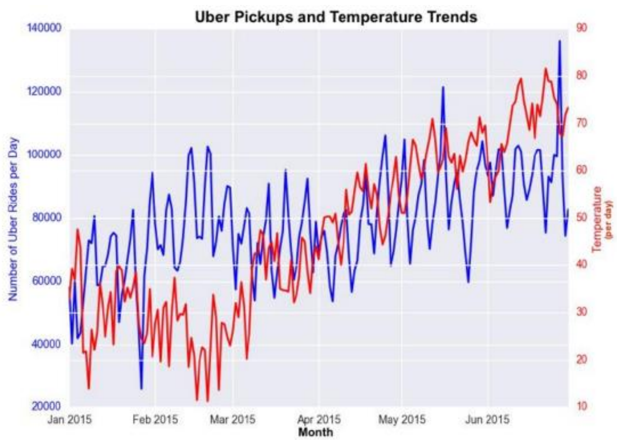


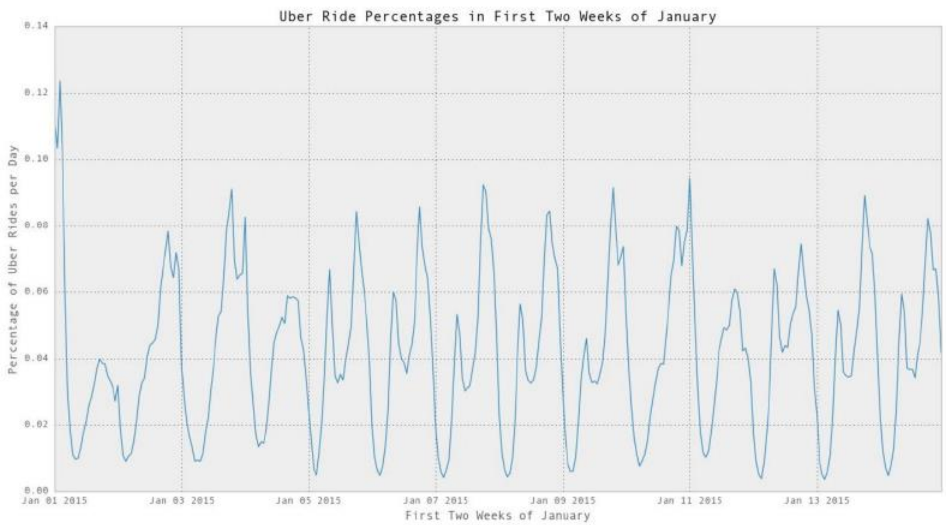


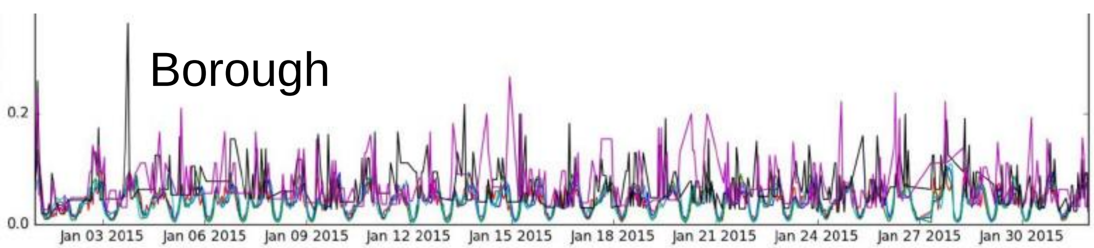
Daphne Angeline Gunawan, Brandon Jihwan Hwang, Alan Yian Xu, Franklin Alexander Velasquez

Uber pickups

NYC Uber Dataset (14.2 million samples)





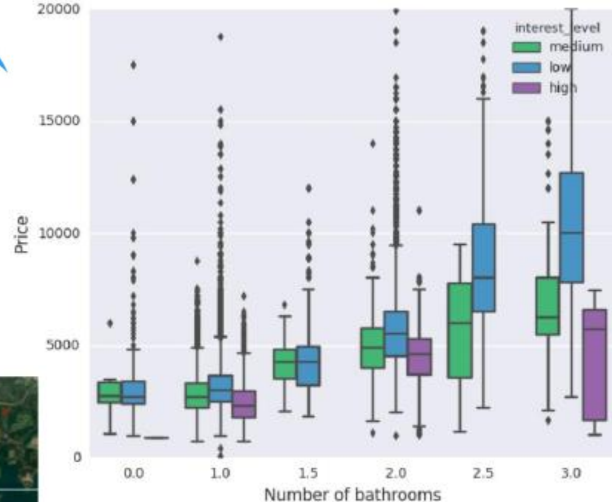


Sample Size	Baseline test MSE	Lstsq train MSE (w/ weather)	Lstsq. Test MSE (w/ weather)	Lstsq train MSE (w/o weather)	Lstsq. Test MSE (w/o weather)
500,000	1.90e-4	1.823e-4	1.189e-3	1.846e-4	1.005e-3
750,000	2.10 e-4	1.874e-4	1.189e-3	1.883e-4	2.070e-4
1000000	4.33 e-4	1.939e-4	1.302e-3	1.942e-4	2.204e-4

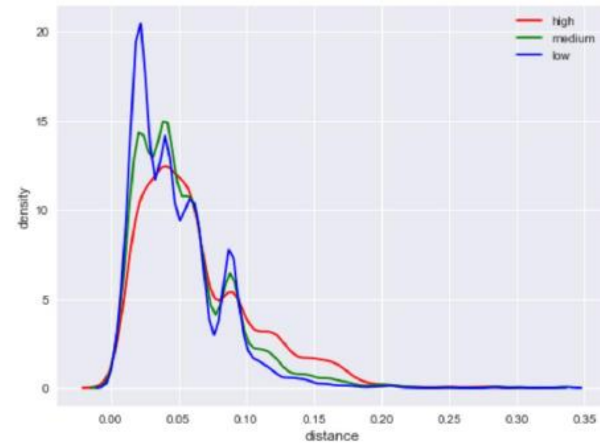
Lilith Huang, Aamir Abdur Rasheed

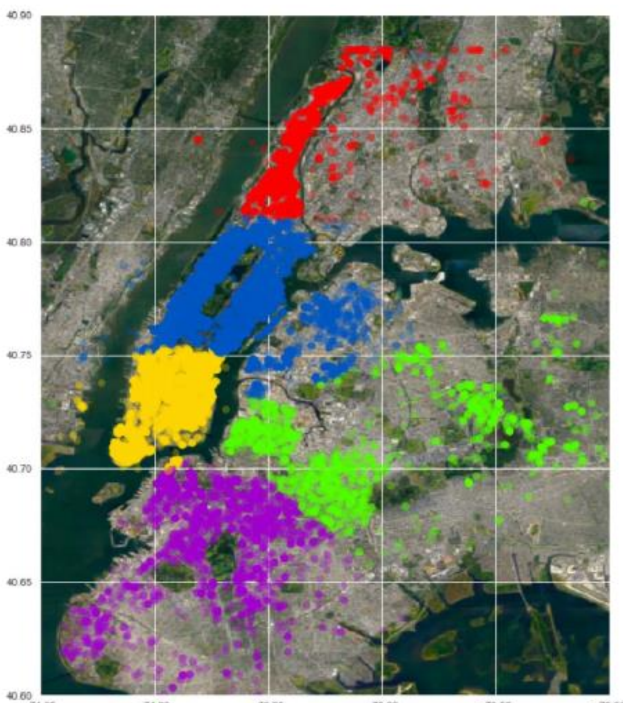
Rental recommendations

#bathrooms



distance to city center



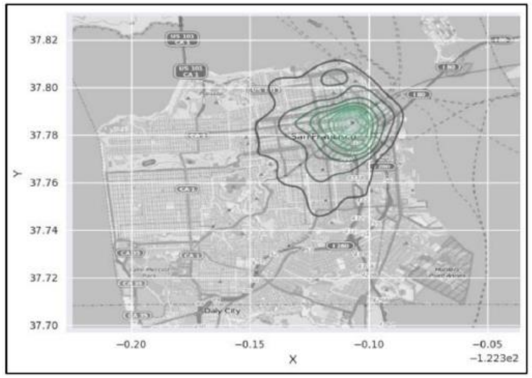


Interest level:

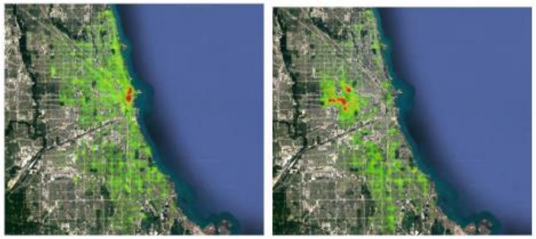
Wen Zhang, Xingbo Wang, Kaixiang Zhao, Lifan Chen
Shiunn An Lu, Shanyu Chuang, Hao-En Sung
Side Li, Yifan Xu
Dhruv Sharma, Keshav Sharma, Saransh Jain

Crime prediction

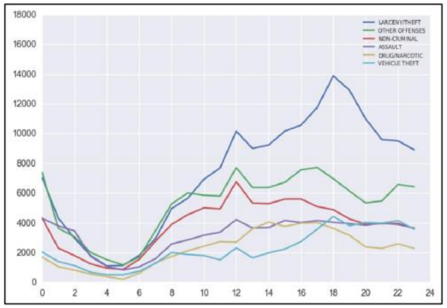
Field	Description
ID	Unique identifier for the record
Case Number	The Chicago Police Department RD Number
Date	Date when the incident occurred in mm/dd/yyyy.
Block	The partially redacted address where the incident occurred.
IUCR	The Illinois Uniform Crime Reporting code
Primary Type	The primary description of the IUCR code.
Description	The secondary description of the IUCR code
Location Description	Description of the location where the incident occurred.
Arrest	Indicates whether an arrest was made.
Domestic	Indicates whether the incident was domestic-related
Beat	Indicates the beat (the smallest police geographic area) where the incident occurred.
District	Indicates the police district where the incident occurred.
Ward	The ward (City Council district) where the incident occurred.
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas.
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
Year	Year the incident occurred.
Updated On	Date and time the record was last updated.
Latitude	The latitude of the location where the incident occurred.
Longitude	The longitude of the location where the incident occurred.



Theft by location

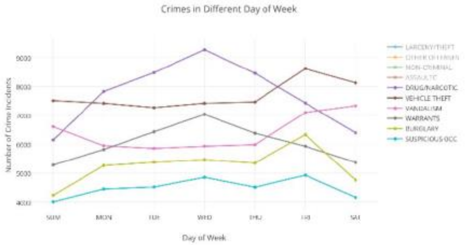


(a) Thefts (b) Narcotics

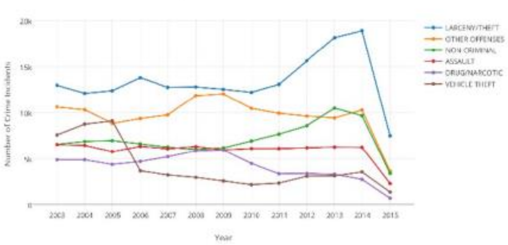


Crime types by hour

Day



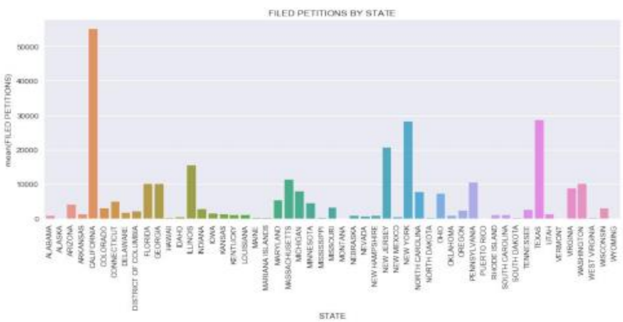
Year



Wenbin Zhu, Yuchen Wang, Wenjie Tao
Sahil Agarwal, Ujjwal Gulecha, Shalini Kedlaya
Junyang Li, Shenghong Wang

H1B petitions

Kaggle dataset (~1 million samples)



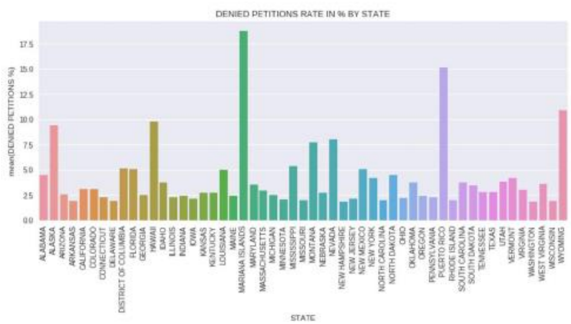


Fig. 8. State-wise Median wage

Company

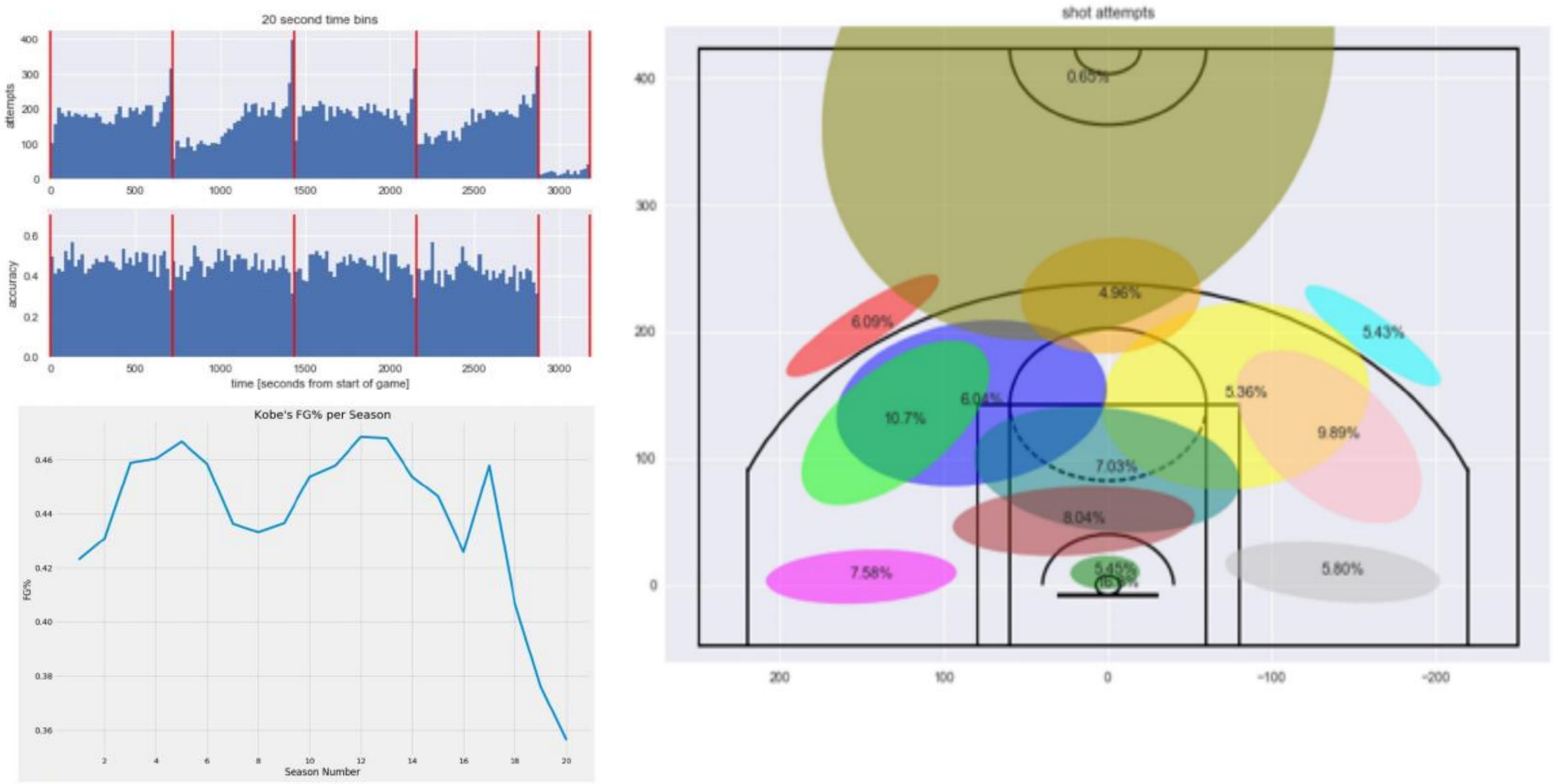
Job title

	SVM	Linear Regression	Logistic Regression
MSE(Training Set)	N\A	0.435	0.393
MSE(Validation set)	N\A	0.527	0.438
MSE(Testing set)	N\A	0.583	0.526
ER(Training set)	0.099	0.059	0.042
ER(Validation Set)	0.125	0.114	0.088
ER(Testing Set)	0.224	0.203	0.127

Yuchen Feng, Xuanzhen Xu, Jianxiong Lin
Prahal Arora, Rahul Vijay Dubey, Induja Sreekanthan, Jahnavi Singhal
Jialin Wang, Yishu Ma, Han Li

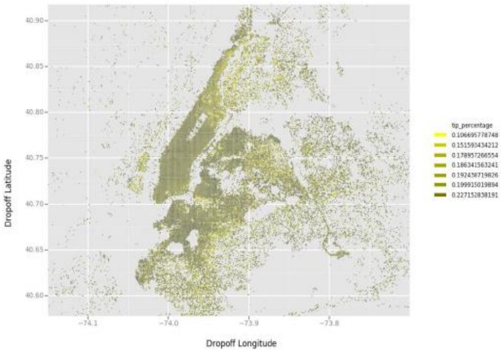
Kobe field goals

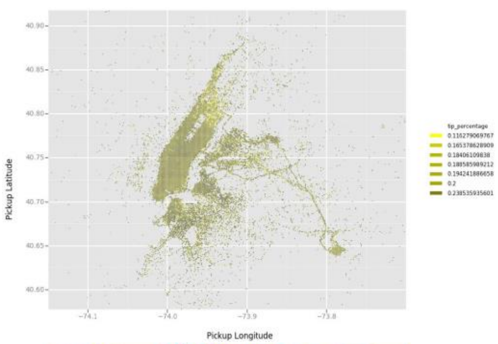
Kaggle competition of 30,000 field-goal attempts



Vishaal Prasad

Taxi tips





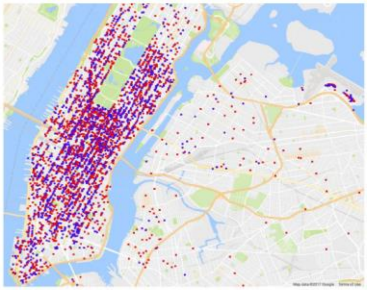
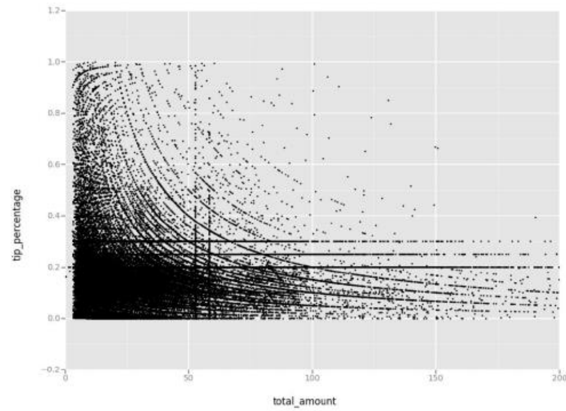
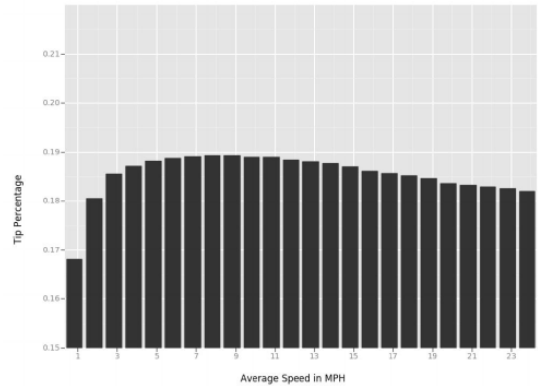
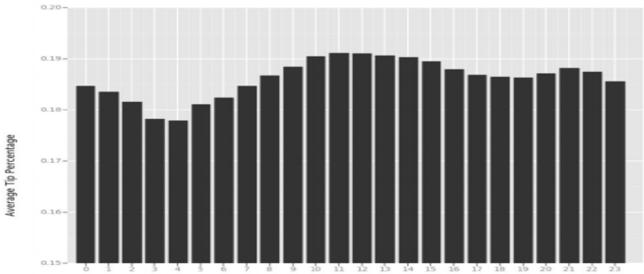
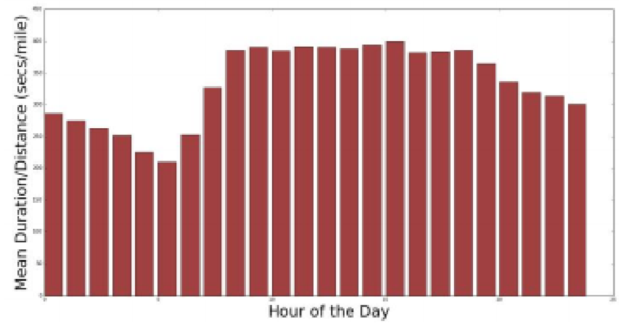


Fig. 1: Mapping of pick-up and drop-off locations









Rushil Nagda, Sudhanshu Bahety, Shubham Gupta
Tejas Saxena, Himanshu Jaiswal, Tushar Bansal, Prateek Ravindra Jakate

Fill out those evaluations!

- Please evaluate the course on https://academicaffairs.ucsd.edu/Modules/Evals?e2250306 !

Thanks!