CSE 158 – Lecture 18

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

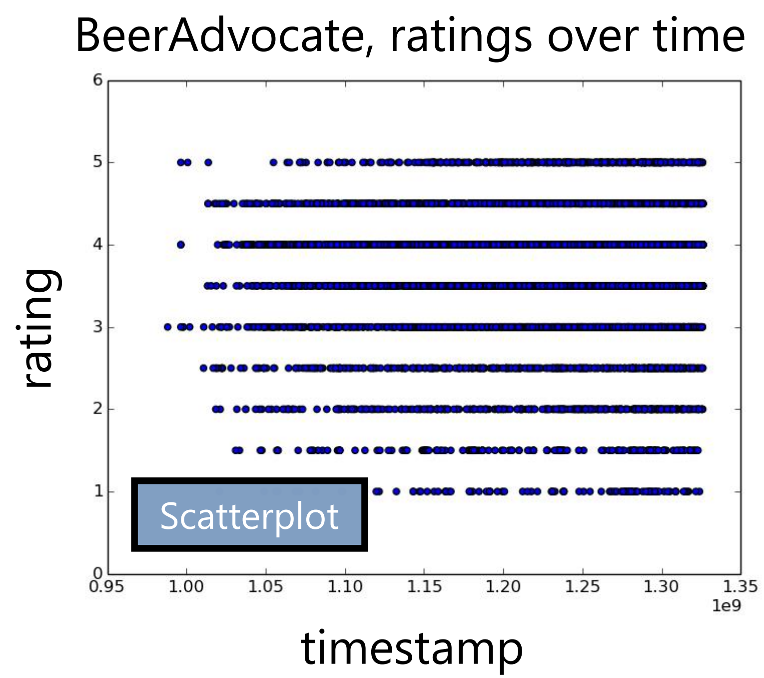
1. **Regression** – sliding windows and autoregression
2. **Classification** – dynamic time-warping
3. **Dimensionality reduction** - ?
4. **Recommender systems** – some results from Koren

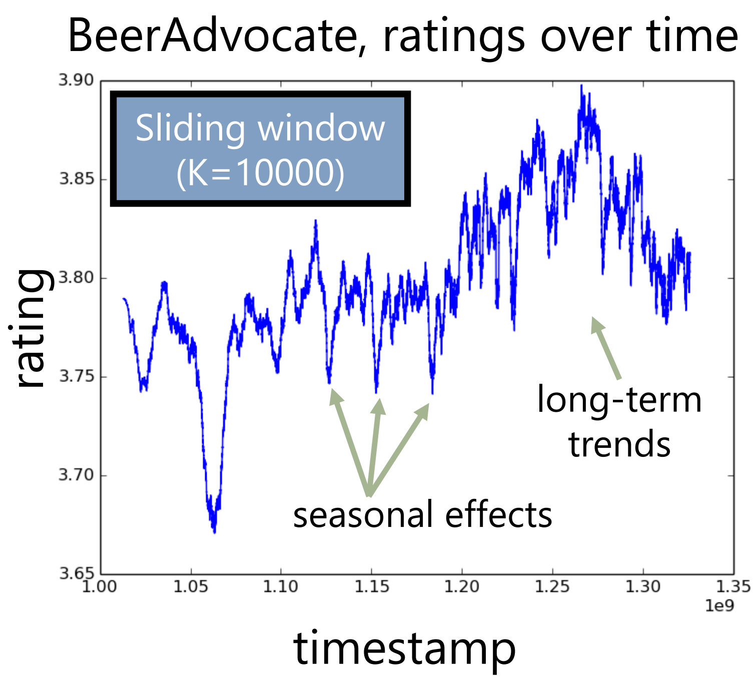
Today:

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/cse258/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

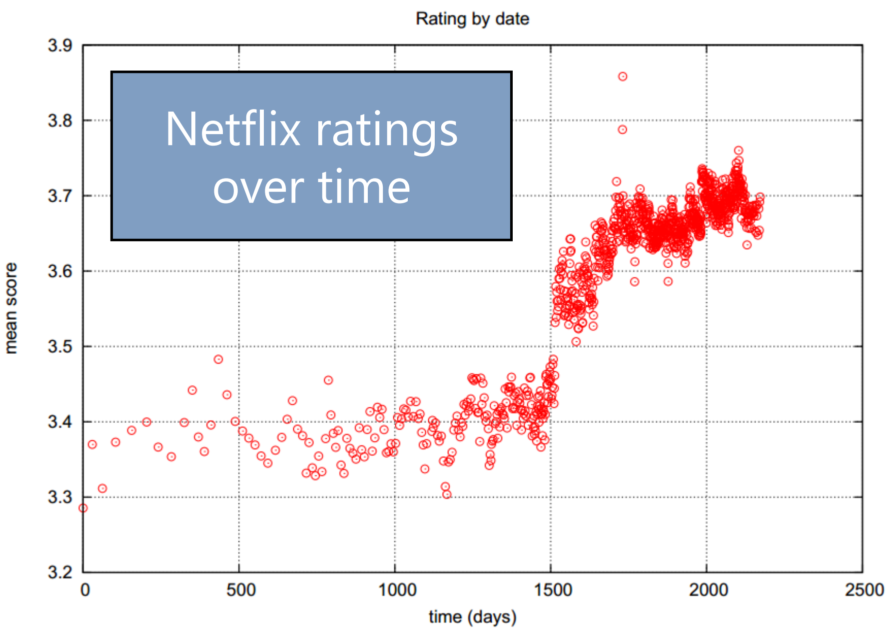
↑ = optimal move is to delete from 2nd sequence

←↑ = either deletion is equally optimal

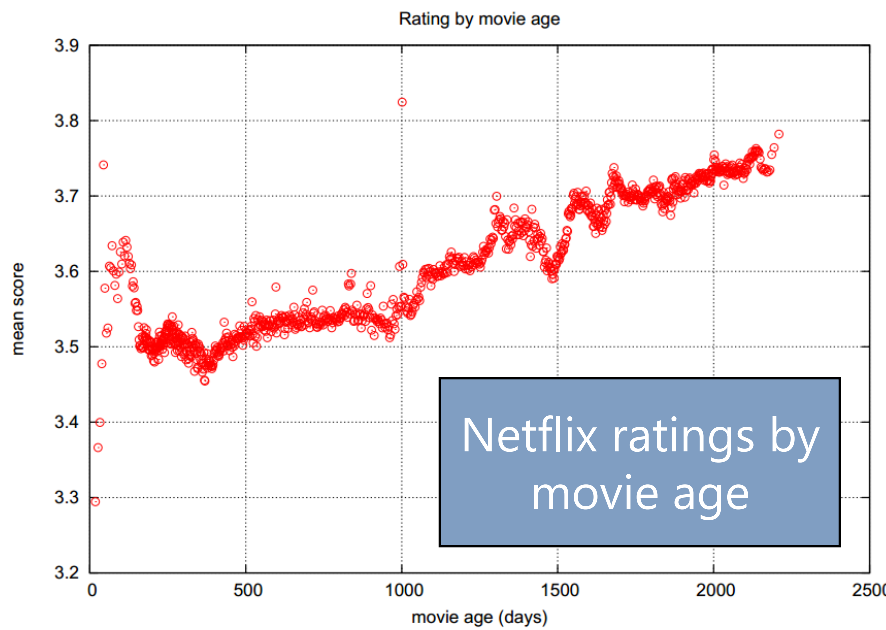
↖ = optimal move is a match

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)



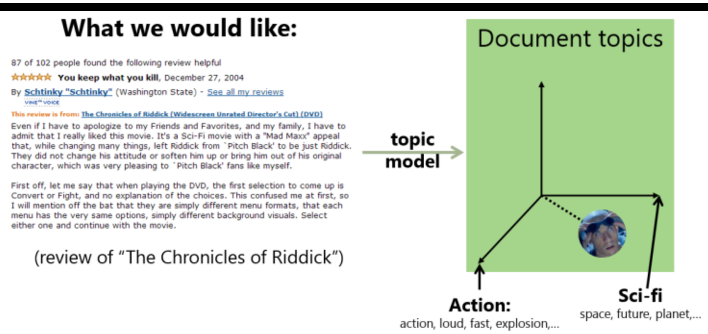
(People tend to give higher ratings to older movies)

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

Week 5: 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

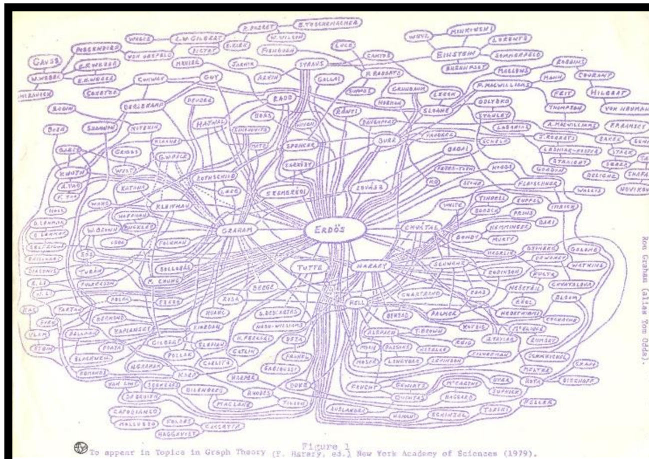
Bags-of-Words



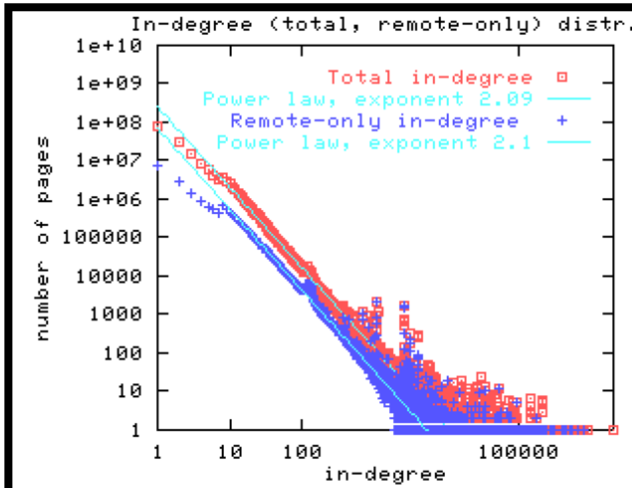
Topic models

Sentiment analysis

8. Social networks

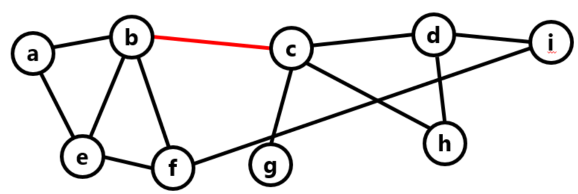


Hubs & authorities



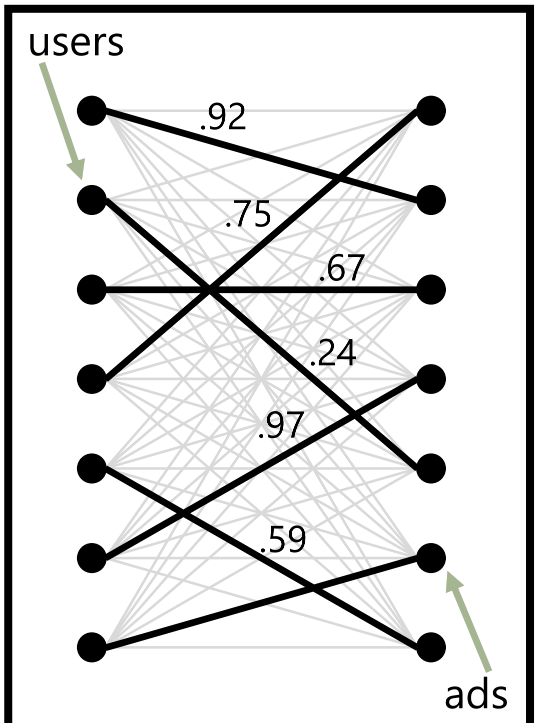
Power laws

Small-world phenomena



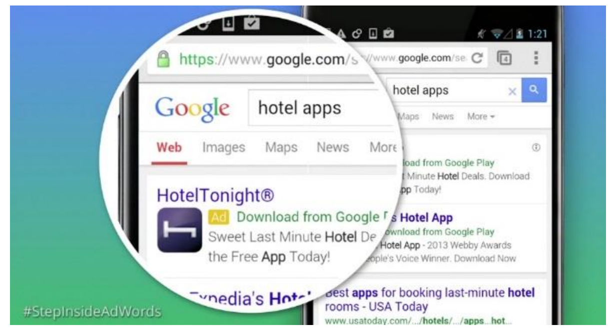
Strong & weak ties

9. Advertising



Matching problems

AdWords



Bandit algorithms

CSE 158 – Lecture 18

Web Mining and Recommender Systems

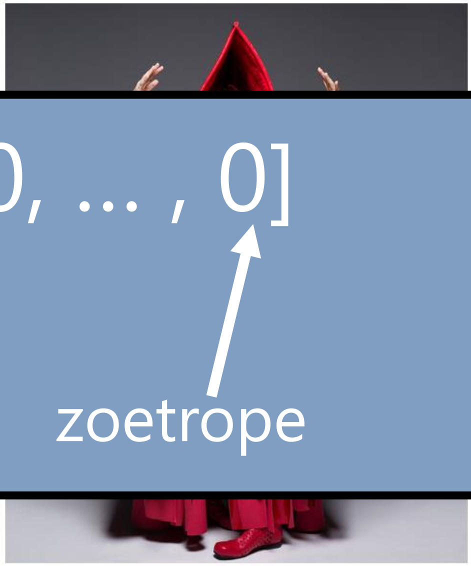
Temporal dynamics of text

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

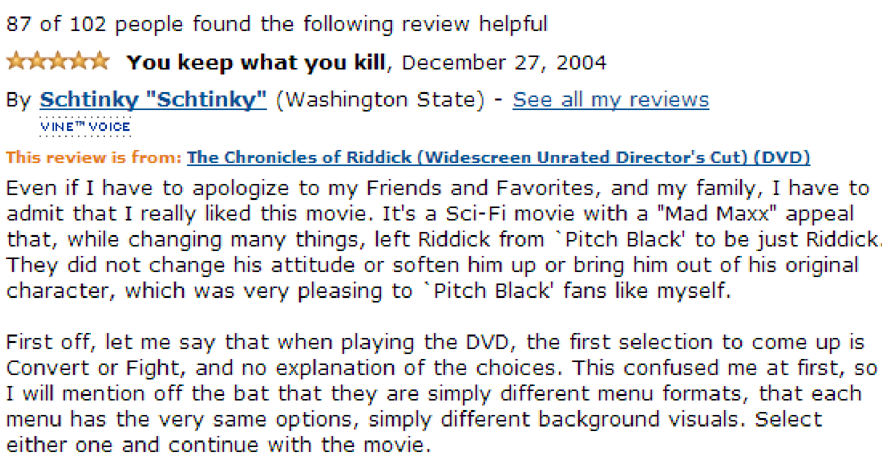
zoetrope

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, 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

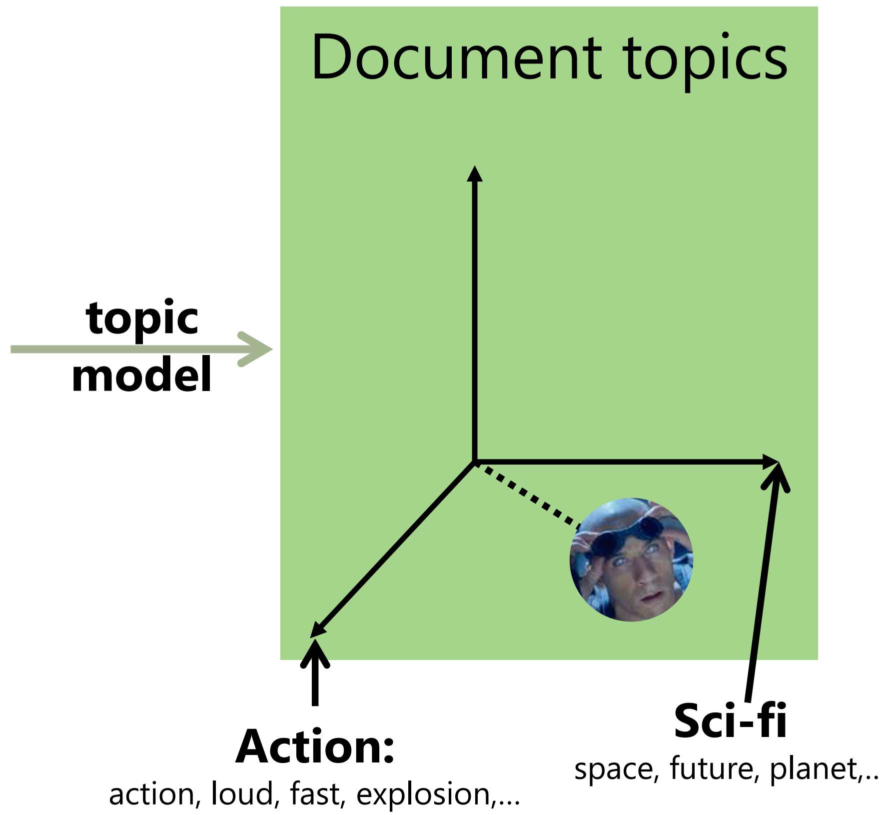
VINE™ VOICE

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.

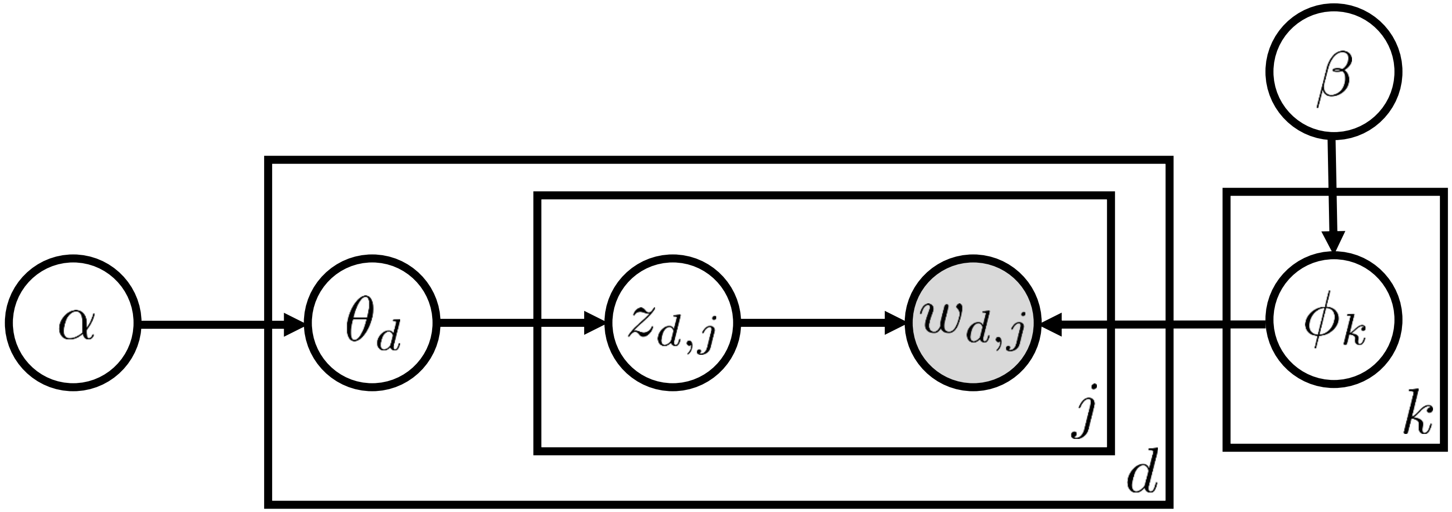
First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background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

(review of "The Chronicles of Riddick")



Latent Dirichlet Allocation

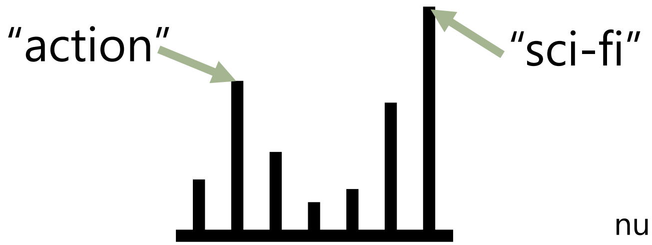
We saw how **LDA** can be used to describe documents in terms of **topics**



- 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:

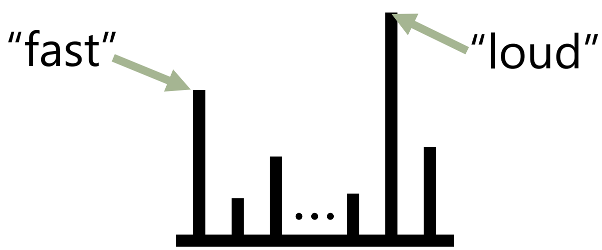


$\theta_{\text{pitch black}}$

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}})**

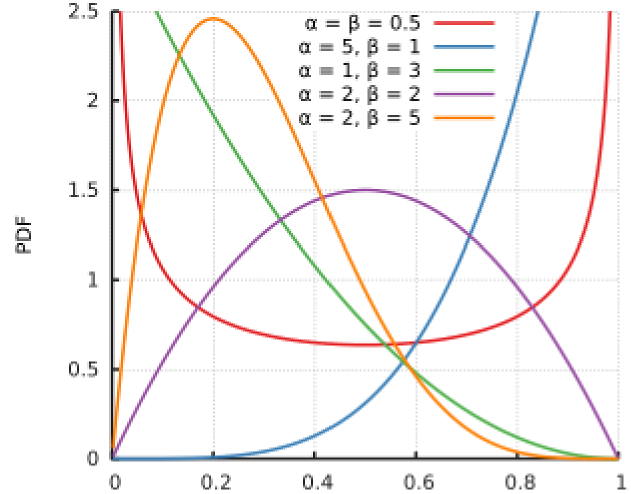
Latent Dirichlet Allocation

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"



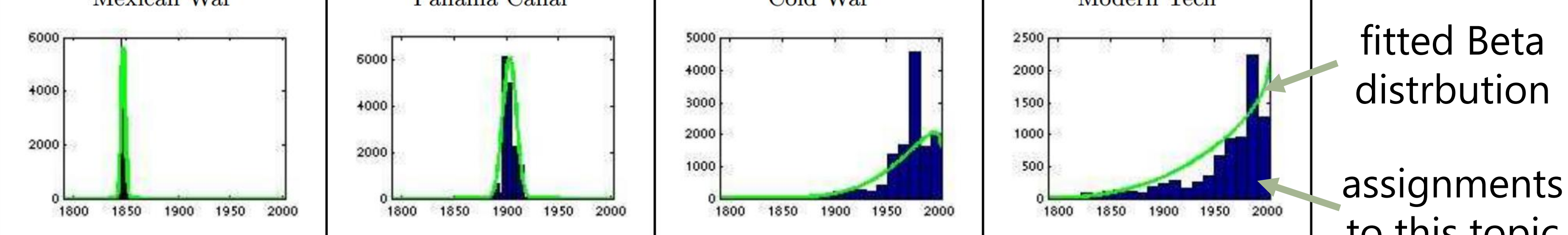
p.d.f.:

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

Latent Dirichlet Allocation

Results:

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

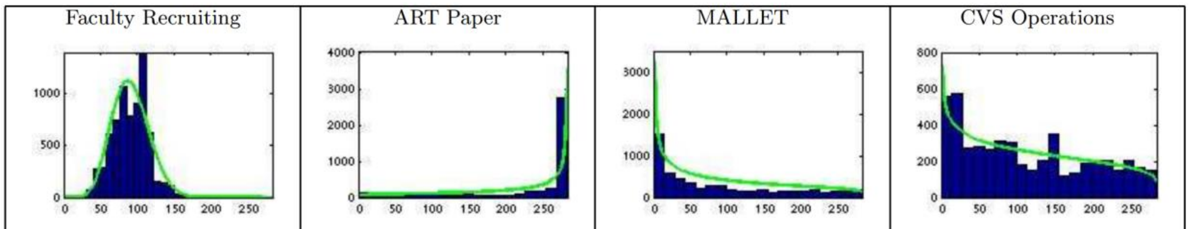


Mexican War		Panama Canal		Cold War		Modern Tech	
states	0.02032	government	0.02928	world	0.01875	energy	0.03902
mexico	0.01832	united	0.02132	states	0.01717	national	0.01534
government	0.01670	states	0.02067	security	0.01710	development	0.01448
united	0.01521	islands	0.01167	soviet	0.01664	space	0.01436
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

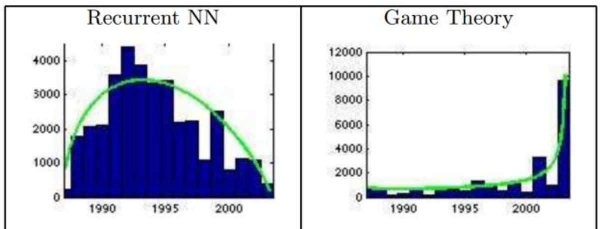
Latent Dirichlet Allocation

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

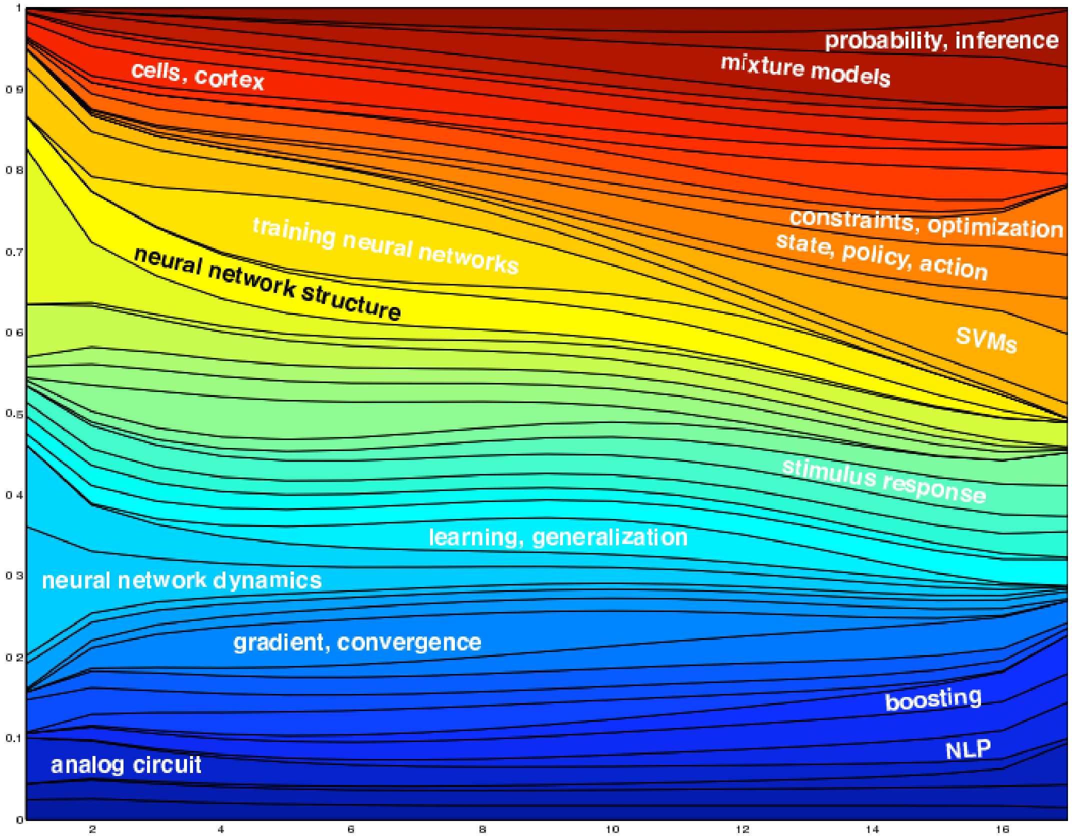


Recurrent NN		Game Theory	
state	0.05963	game	0.02850
recurrent	0.03765	strategy	0.02378
sequence	0.03616	play	0.01490
sequences	0.02462	games	0.01473
time	0.02402	player	0.01451
states	0.02057	agents	0.01346
transition	0.01300	expert	0.01281
finite	0.01242	strategies	0.01123
length	0.01154	opponent...	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 158 – Lecture 18

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

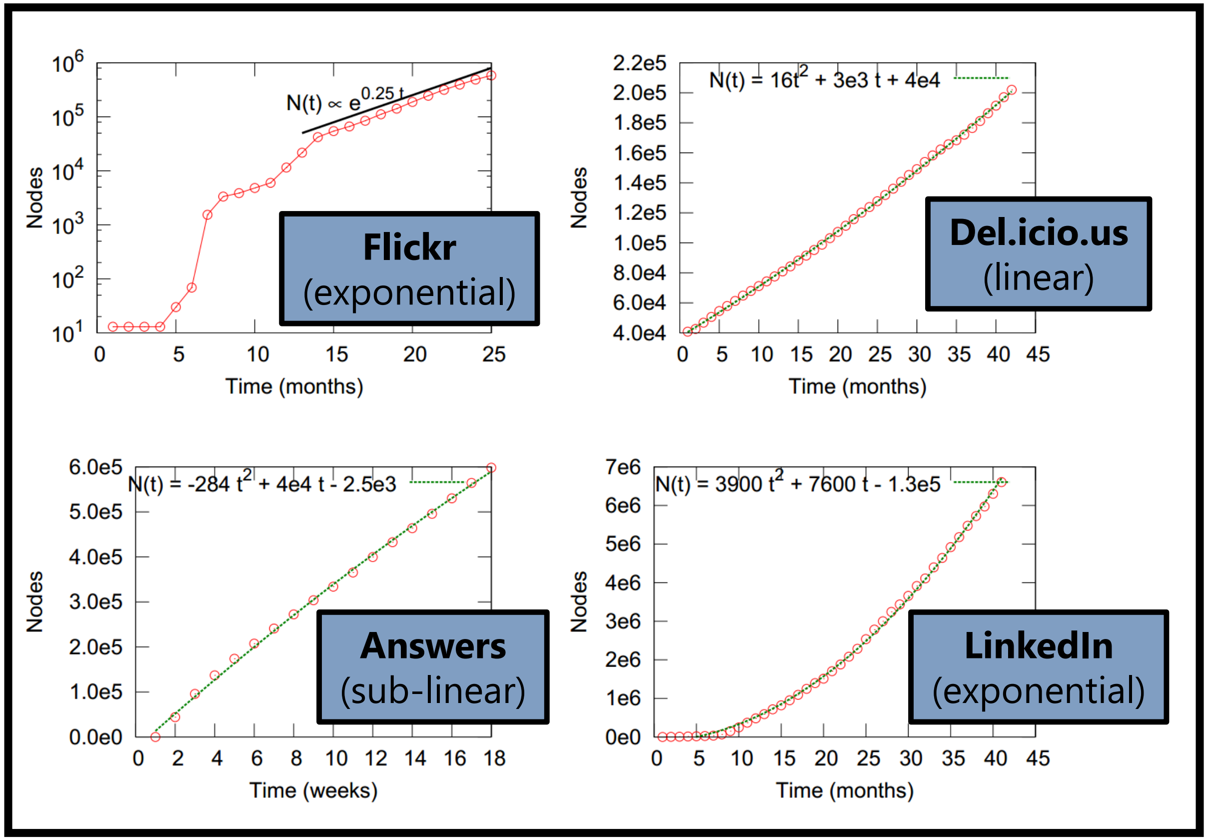
- Power-laws & small worlds
- Random graph models

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?



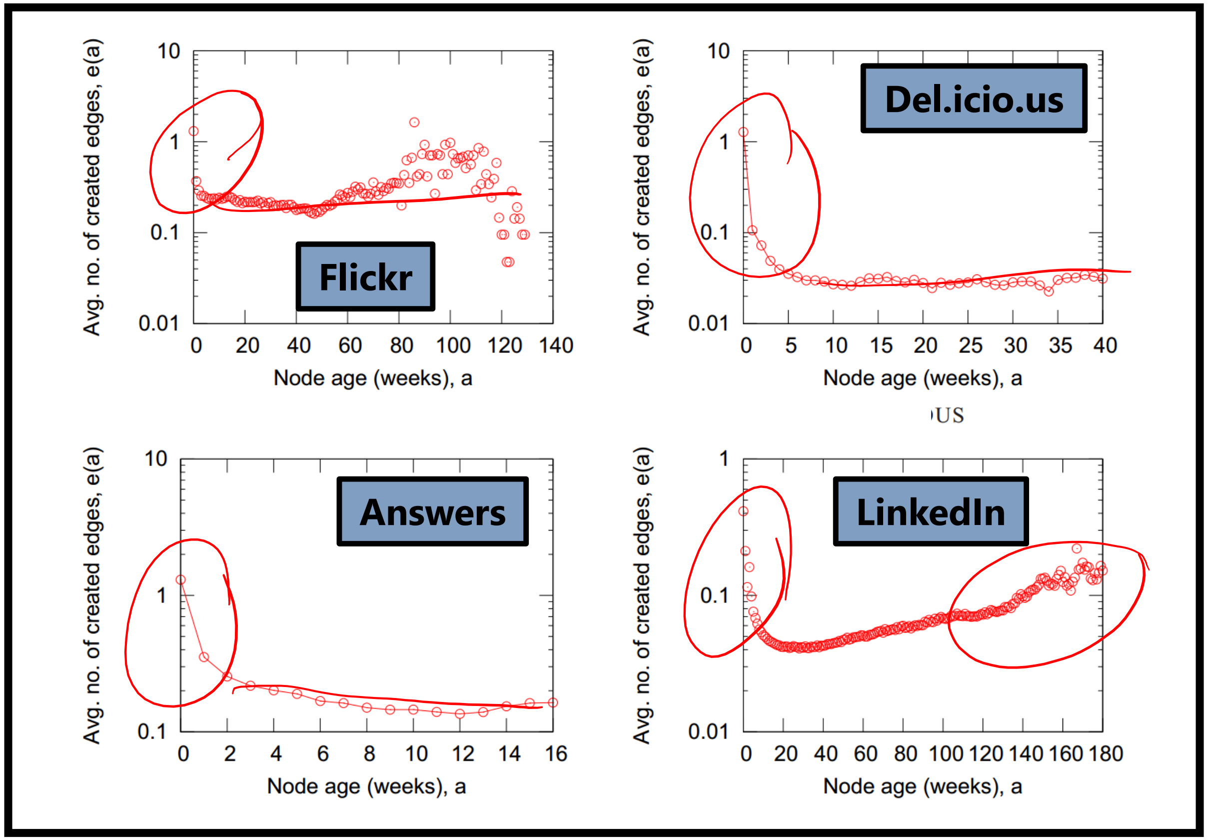
(from Leskovec, 2008 (CMU Thesis))

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

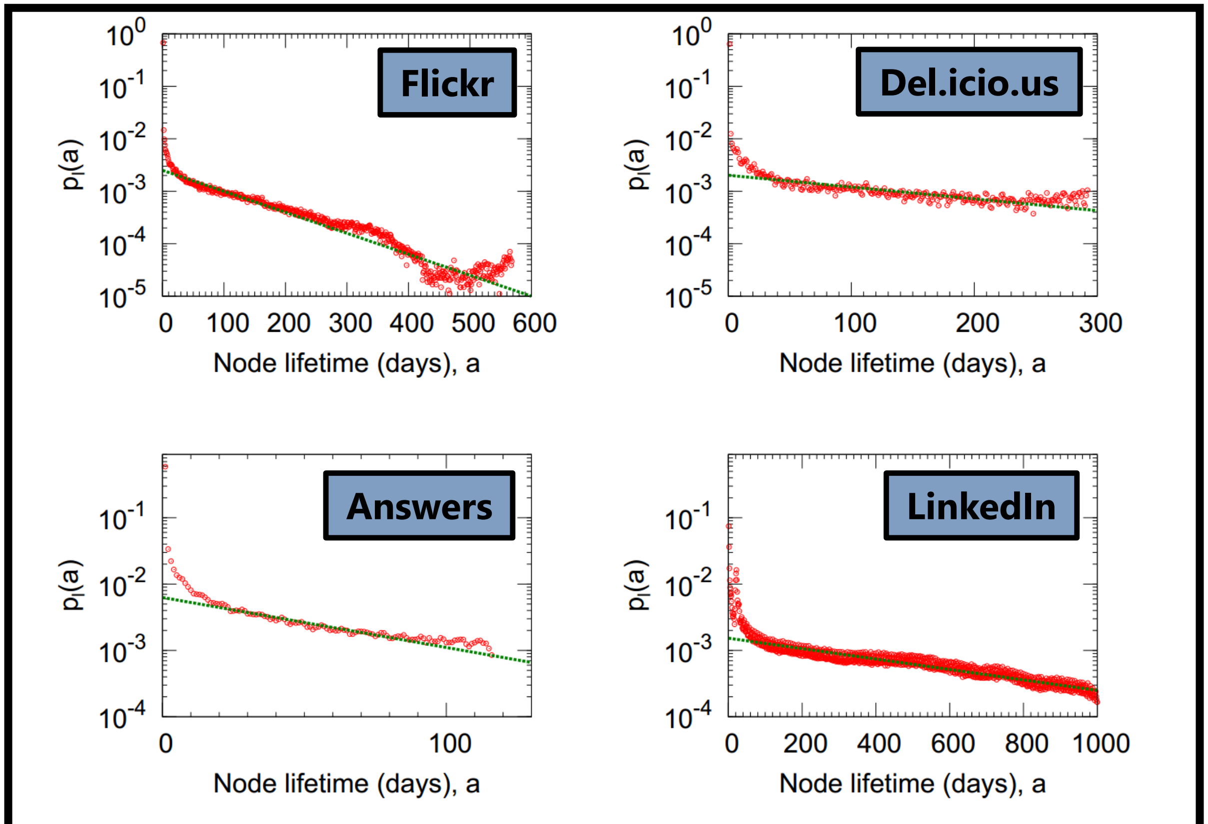


A: In most networks there's a "burst" of initial edge creation which gradually flattens out.
Very different behavior on LinkedIn (guesses as to why?)

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



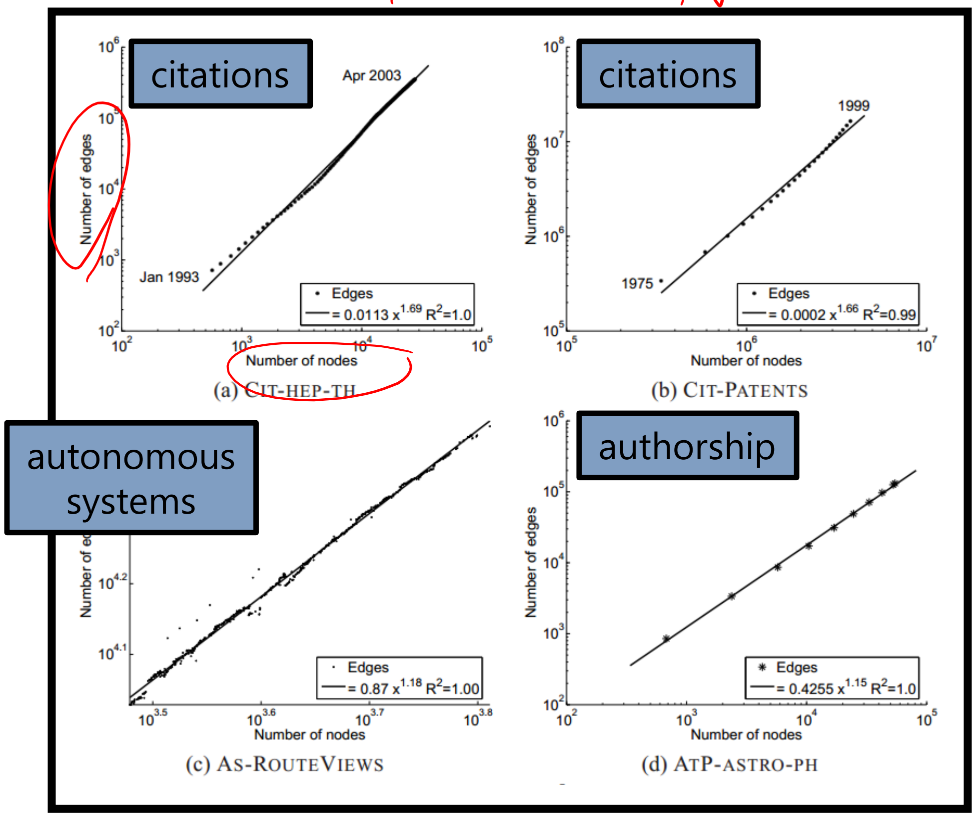
A: Node lifetimes follow a power-law: many many nodes are shortlived, with a long-tail of older nodes

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?

$\#E = cN^{\alpha} \quad \alpha > 1$



- 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

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

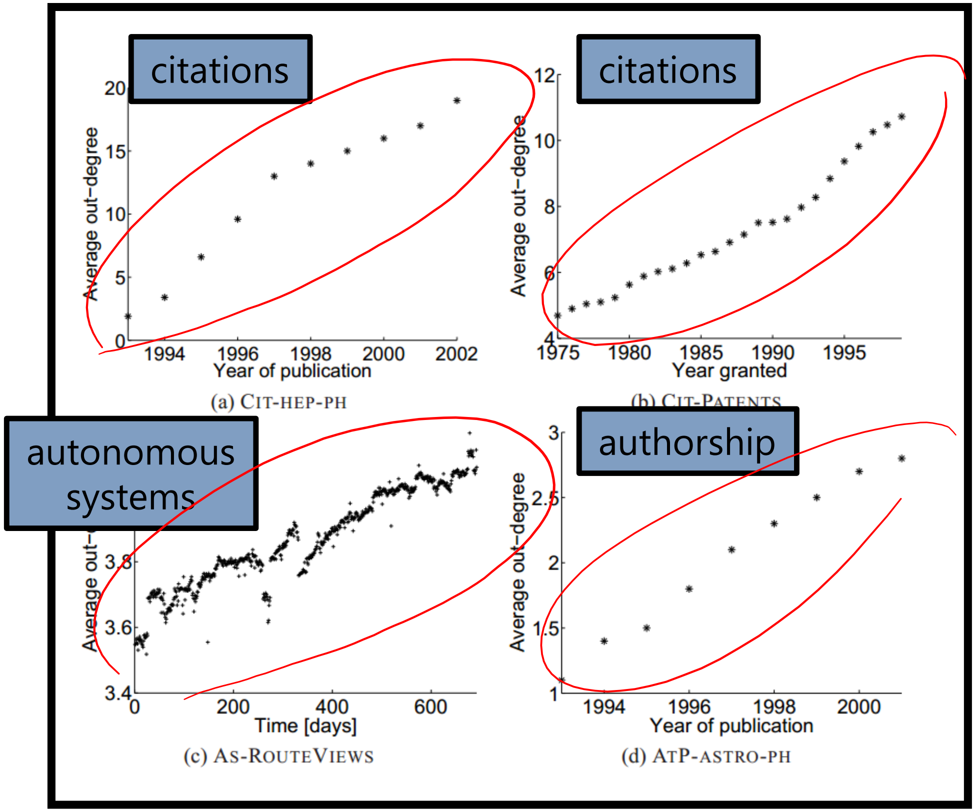
where

$$1 \leq a \leq 2$$

- 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

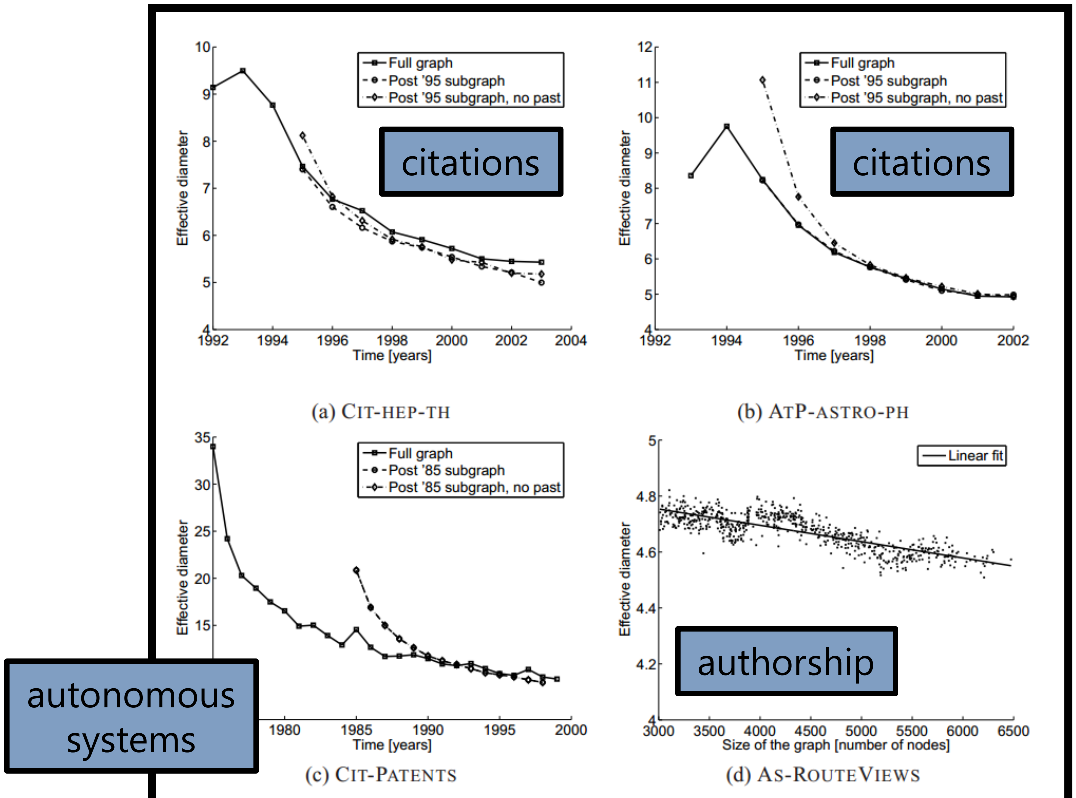
Q2: How does the degree change over time?



- **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

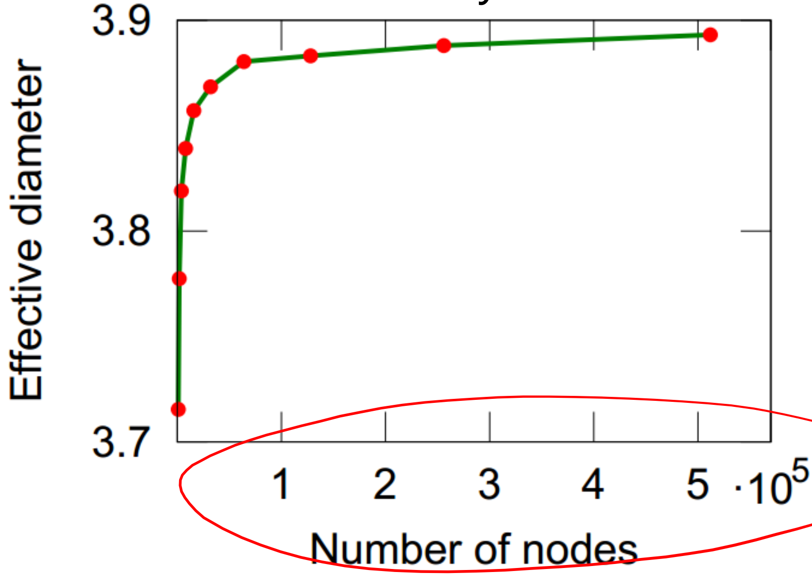
Temporal dynamics of social networks

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:

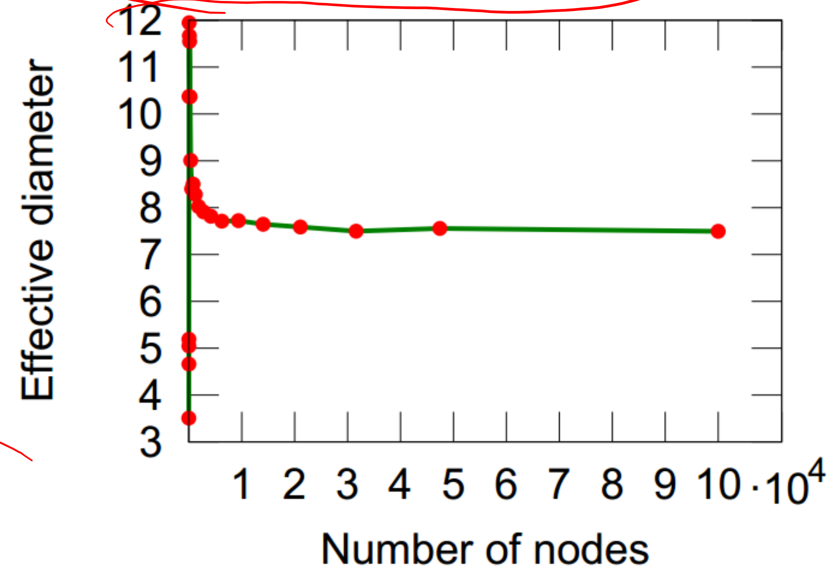
$E = N^{1.3}$

Erdos-Renyi – a = 1.3



$E = N^{1.2}$

Pref. attachment model – a = 1.2

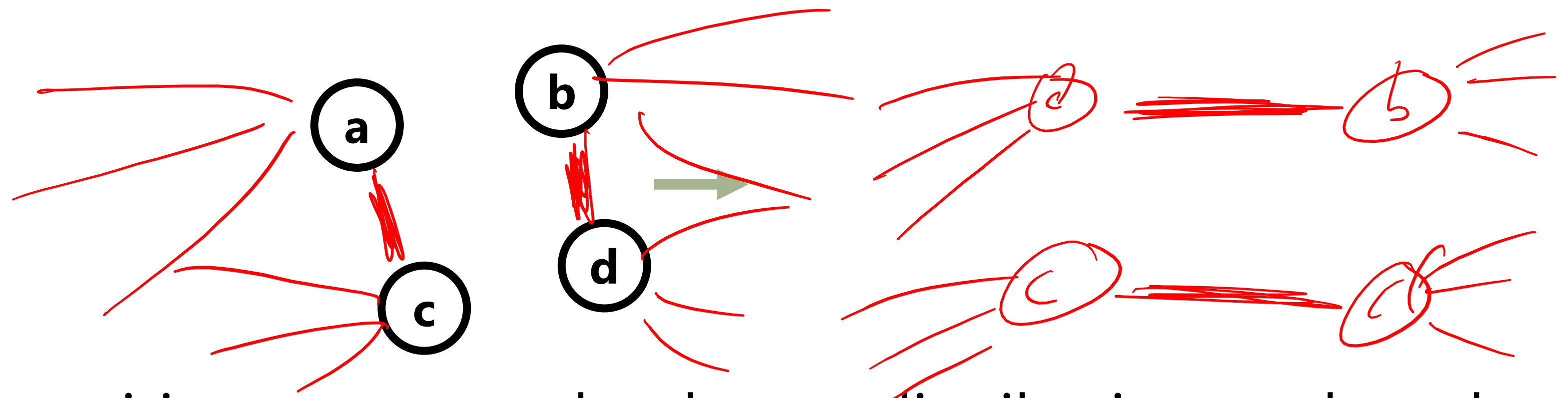


Temporal dynamics of social networks

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

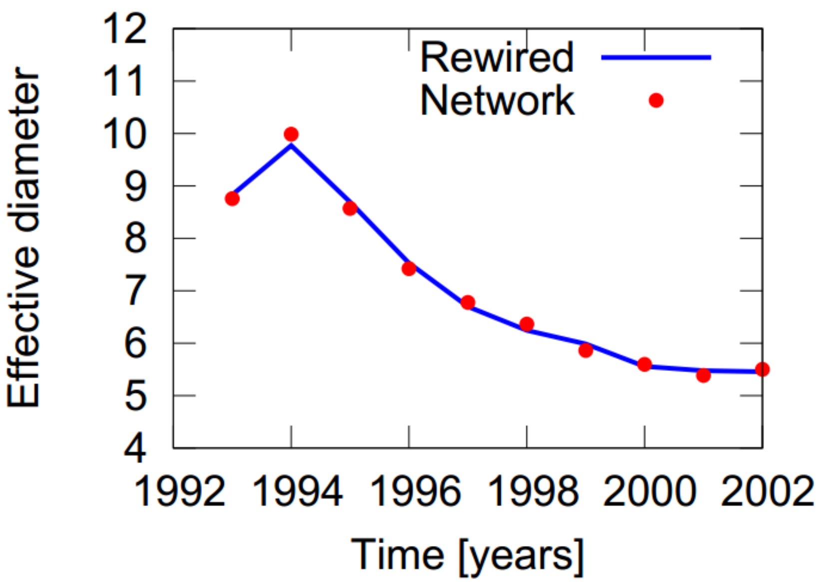


random rewiring preserves the degree distribution, and randomly samples amongst networks with observed degree distribution

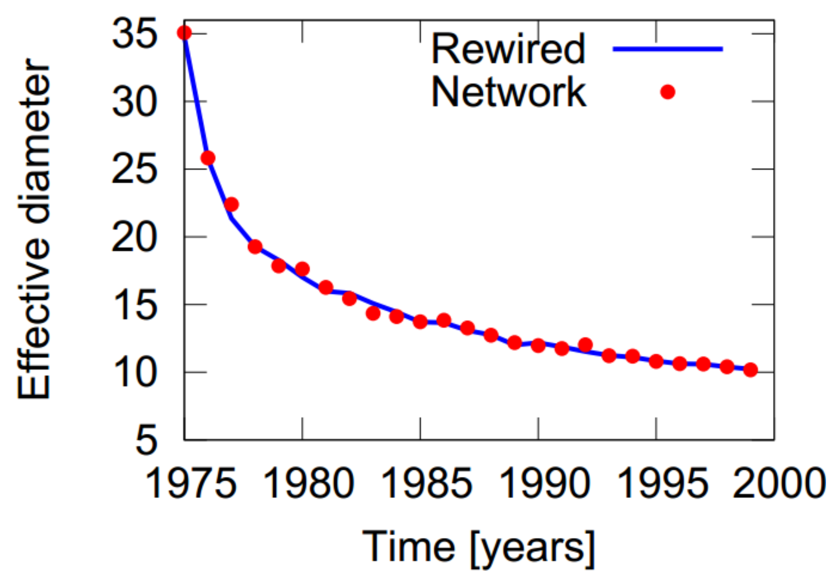
Temporal dynamics of social networks

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

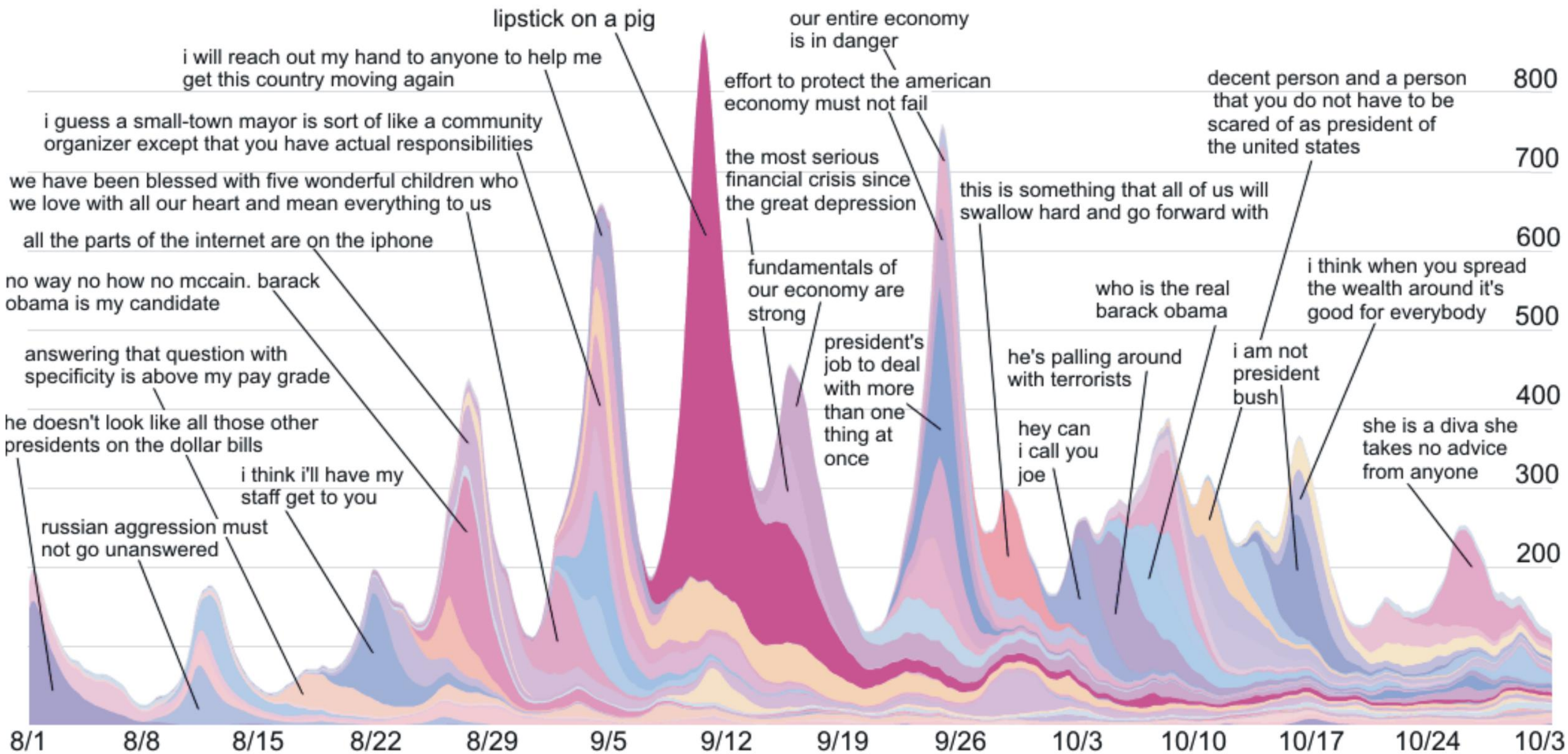
So, a decreasing diameter is **not** a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model

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

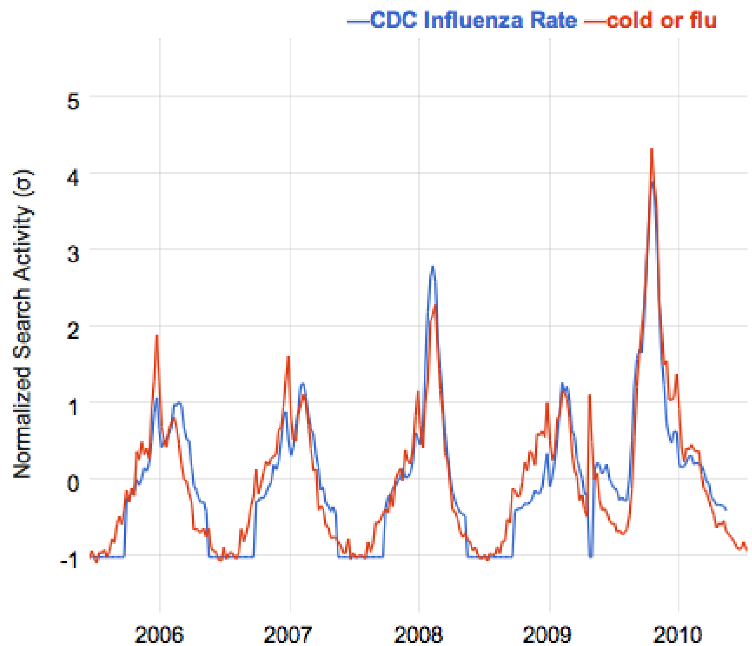
Other interesting topics…



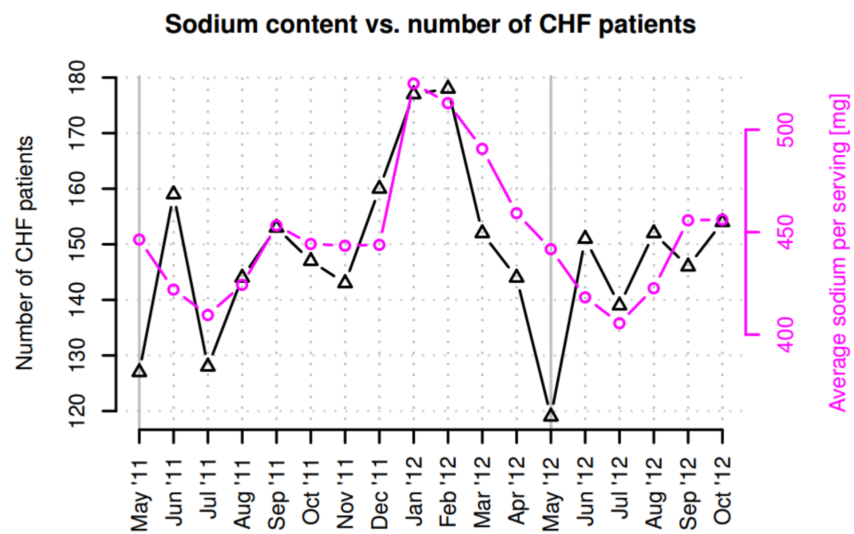
"memetracker"

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

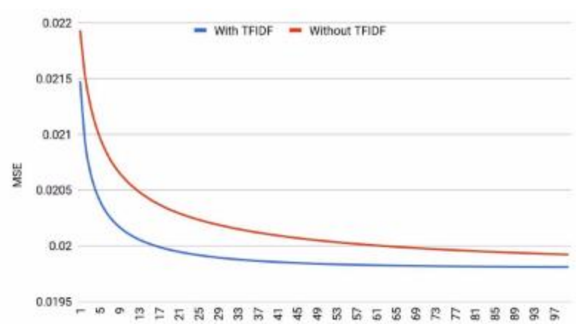
CSE 158 – Lecture 18

Web Mining and Recommender Systems

Some incredible assignments

Supervised funniness detection in the New Yorker cartoon caption contest

"I was just transferred to the fraternity ward."



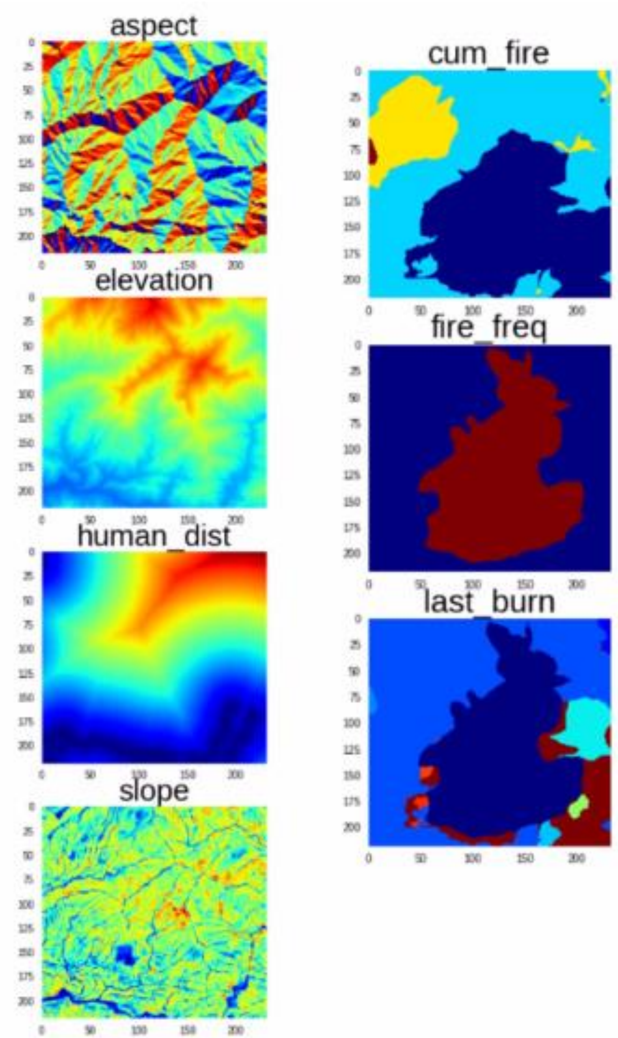
TF-IDF vs non-TF-IDF models

- Predict whether a caption will be scored as “funny” by human judges
- 65 images, 320k captions
- Scores from 1.0 – 2.75

- BoW methods w/ and w/o TF-IDF
- Dimensionality-reduction-based feature representations

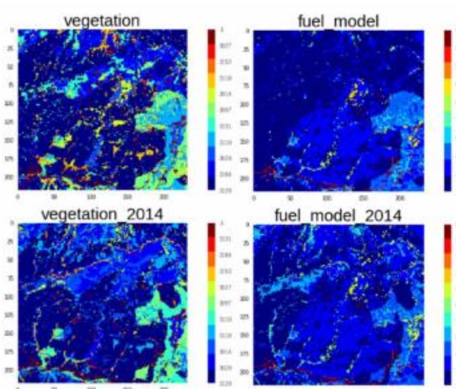
Melissa Wright

Predicting Vegetation Changes as Responses to Forest Fires



- Geological data from LANDFIRE program and FRAP (Fire and Resource Assessment Program), 1992-2012
- Estimate changes as a result of forest fires

$$y = x_{2012\ vegetation} == x_{2014\ vegetation}\ \forall x \in X$$



	0
human_dist	0.214184
elevation	0.163118
vegetation	0.123
aspect	0.0872108
slope	0.0770156
VEG_3986	0.0517486
cum_fire	0.041889
fuel_model	0.0265596
VEG_3008	0.0256087
VEG_3221	0.0159329

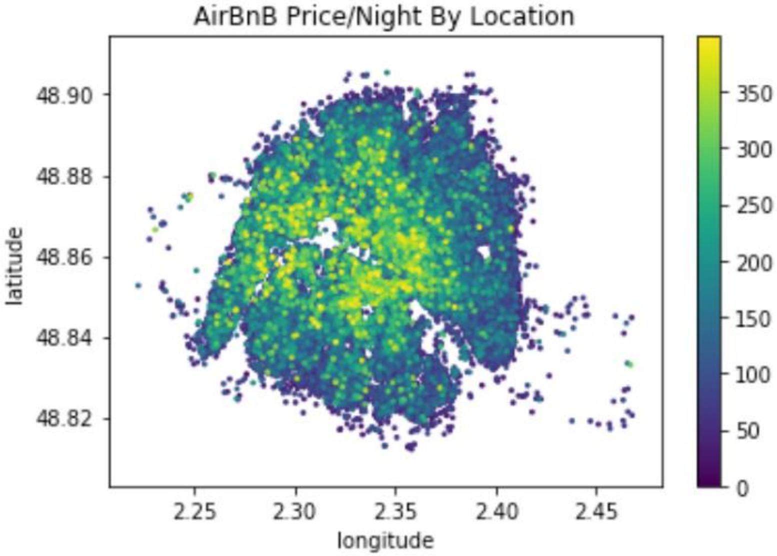
Feature importance from Random Forest Model

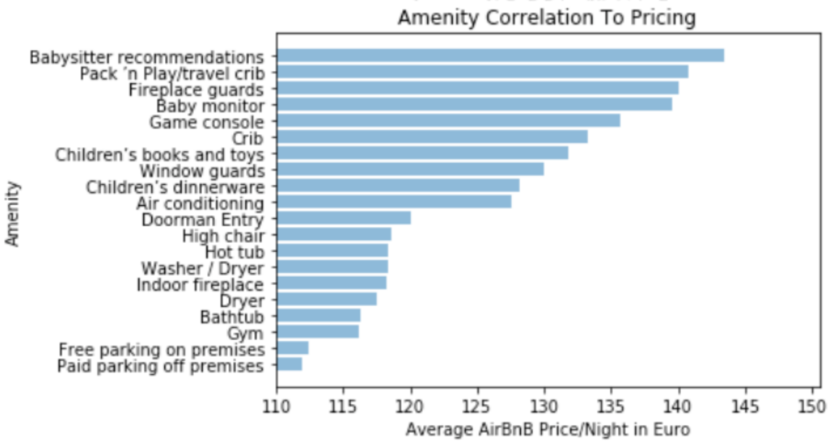
Tony Salim

AirBnB Price Per Night Prediction

Price Range	€ 0.00 to € 7,790.00
Mean	€ 96.12
Median	€ 75.00
Standard Deviation	€ 99.30

- AirBnB Paris data
- Predict listing price given various features

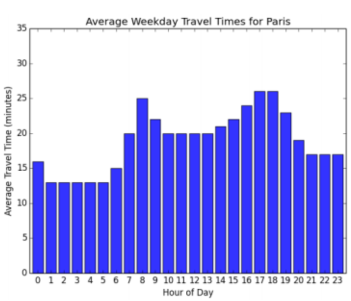


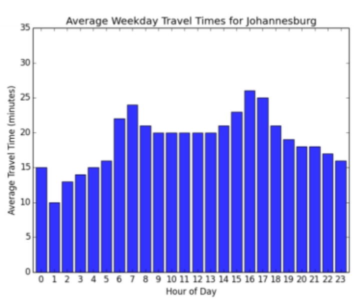


Peter Mai

Uber Everywhere: Exploring Movement

Feature	Description
Hour of day (hod)	Simple hour of the day feature.
Source ID	Simple source ID feature.
Destination ID	Simple destination ID feature.
Hour of day historical mean*	Mean travel category of trips for this hour of day.
Source ID historical mean*	Mean travel category of trips from this source ID.
Destination ID historical mean*	Mean travel category of trips from this destination ID.
Source-Destination ID pair historical mean*	Mean travel category of trips from specific source ID-destination ID pair.





Weekday travel times in two cities

- Anonymized Uber Movement data from 7 cities
- Trip time given source, destination, and hour

Feature Representation	Week Category	Results
hod, source ID, dest ID	Weekday	26.544%
	Weekend	29.247%
hod mean, source ID mean, dest ID mean	Weekday	26.788%
	Weekend	29.113%
hod, source ID, dest ID, hod mean, source ID mean, dest ID mean, combined source ID-dest ID mean	Weekday	21.318%
	Weekend	25.024%
hod, combined source ID-dest ID mean	Weekday	79.218% / 79.975%*
	Weekend	87.041% / 87.146%*

SVM,
Random Forest
MLP

Tynan Dewes, David Thomson

Predicting the Accepted Answer for StackOverflow Questions

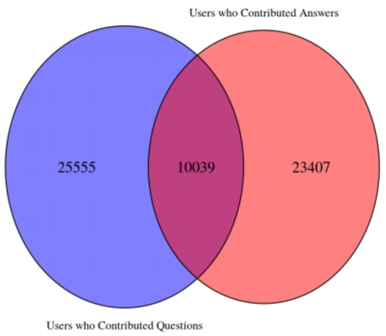
Figure 1: Example Entry in Posts.xml

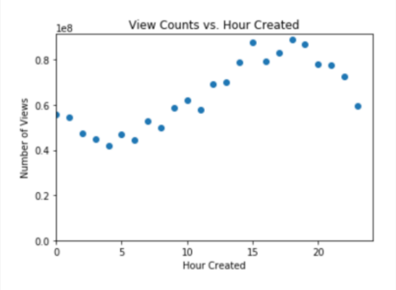
```
<row Id="4" PostTypeId="1"
    AcceptedAnswerId="7" CreationDate="2008-07-31T21:42:52.667" Score="506"
    ViewCount="32399" Body="&lt;p&gt;I want to use a track-bar to change a
    form's opacity.&lt;/p&gt;&#xA;&#xA;&lt;p&gt;This is my
    code:&lt;/p&gt;&#xA;&#xA;&lt;pre&gt;&lt;code&gt;decimal trans =
    trackBar1.Value / 5000;&#xA;this.Opacity =
    trans;&#xA;&lt;/code&gt;&lt;/pre&gt;&#xA;&#xA;&lt;p&gt;When I build the
    application, it gives the following
    error:&lt;/p&gt;&#xA;&#xA;&lt;blockquote&gt;&#xA;  &lt;p&gt;Cannot
    implicitly convert type 'decimal' to
    'double'.&lt;/p&gt;&#xA;&lt;/blockquote&gt;&#xA;&#xA;&lt;p&gt;I tried
    using &lt;code&gt;trans&lt;/code&gt; and &lt;code&gt;double&lt;/code&gt;
    but then the control doesn't work. This code worked fine in a past
    VB.NET project. &lt;/p&gt;&#xA;" OwnerUserId="8"
    LastEditorUserId="126970" LastEditorDisplayName="Rich B"
    LastEditDate="2017-03-10T15:18:33.147"
    LastActivityDate="2017-03-10T15:18:33.147" Title="While applying opacity
    to a form should we use a decimal or double value?"
    Tags="&lt;c#&gt;&lt;winforms&gt;&lt;type-conversion&gt;&lt;decimal&gt;&lt;opacity&gt;"
    AnswerCount="13" CommentCount="5" FavoriteCount="37"
    CommunityOwnedDate="2012-10-31T16:42:47.213" />
```

```
{ int(QuestionID):  {   "q":           {
                                  'AcceptedAnswerId':int(ID of answer that was accepted by poster),
                                  'AnswerCount':int(number of answers to the question),
                                  'Body': {
                                           'code':[text of code tags],
                                           'links':[text of a.href values],
                                           'text':str(plain text of question)
                                        },
                          [opt]   'ClosedDate':datetime(date when the question was closed),
                                  'CommentCount':int(number of comments on question),
                          [opt]   'CommunityOwnedDate':datetime(date when question was community owned),
                                  'CreationDate':datetime(date question was created),
                          [opt]   'FavoriteCount':int(number of favroties on question),
                                  'Id':int(id of the question),
                                  'LastActivityDate':datetime(date of last activity on the question),
                          [opt]   'LastEditDate':datetime(date when the question was last edited),
                          [opt]   'LastEditorDisplayName':str(name of the last editor on the question),
                          [opt]   'LastEditorUserId':int(id of user who last edited),
                          [opt]   'OwnerDisplayName':str(name of the answer's writer),
                          [opt]   'OwnerUserId':int(user id of the owner),
                                  'PostTypeId':int(1 if question, 2 if answer),
                                  'Score':int(score of the question),
                                  'Tags':[str(tags) on question],
                                  'Title':str(title of question),
                                  'ViewCount':int(number of views on question),
                                  'paragraphs':int(number of newlines in question text)
                               }
foreach question ans>>
                  int(AnswerID):  {
                                  'Body': {
                                           'code':[text of code tags],
                                           'links':[text of a.href values],
                                           'text':str(plain text of answer)
                                        },
                                  'CommentCount':int(number of comments on answer)
                          [opt]   'CommunityOwnedDate':datetime(date when answer was community owned),
                                  'CreationDate':datetime(date answer was created),
                                  'Id':int(id of the answer),
                                  'LastActivityDate':datetime(date of last activity on the answer),
                          [opt]   'LastEditDate':datetime(date of the last edit on the answer),
                          [opt]   'LastEditorDisplayName':str(name of the last editor of the question),
                          [opt]   'LastEditorUserId':int(id of last user who edited answer),
                          [opt]   'OwnerUserId':int(id of the answer's writer),
                          [opt]   'OwnerDisplayName':str(name of the answer's writer),
                                  'ParentId':int(id of the question the answer is for),
                                  'PostTypeId':int(1 if question, 2 if answer),
                                  'Score':int(score of the answer),
                                  'paragraphs':int(number of newlines in answer text)
                                }

                  }
}
```

- Large dataset of StackOverflow posts
- Predict which answer is marked as "accepted" (classification)





Feature	Type
Answer Score	int
Answer Creation Month	int in range(1,13)
Difference in Seconds between Answer Creation and Question Creation	float
Difference in Seconds between Last Answer Activty and Answer Creation	float
Answer Comment Count	int
Percentage of Total Answer Link Count for this Question this Answer Accounts For	float
Percentage of Total Answer Code Entry Count for this Question this Answer Accounts For	float
Number of Words in Answer	int
Total Number of Answers to Question	int
Number of Words in Question Title	int
Number of Views on Question	int
Numer of Paragraphs in Answer	int
Number of Paragraphs in Question	int
Whether or not Answer was Edited	bool
Answer Creation Year	int
Answer Creation Hour	int in range(0,25)

Mustafa Guler, Jessica Kwok, Joseph Thomas

Bitcoin Price Prediction using ARIMA, Linear Regression and Deep Learning

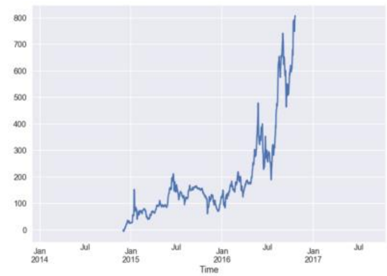


Fig. 4. Percentage Return on Investment in 1 year

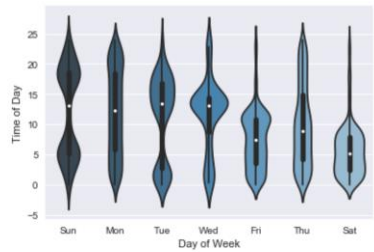
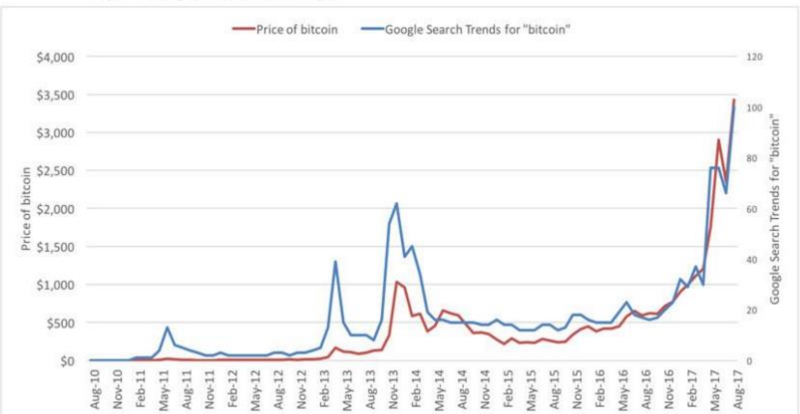


Fig. 6. Violin plot describing best time of day to invest in bitcoin



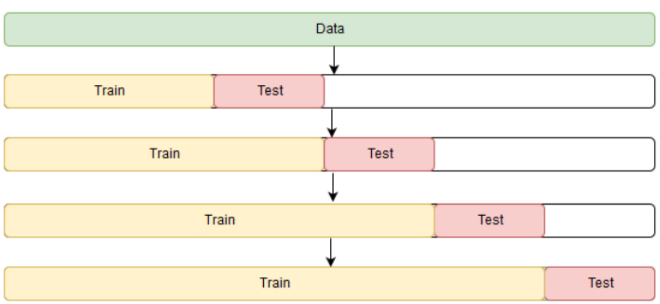
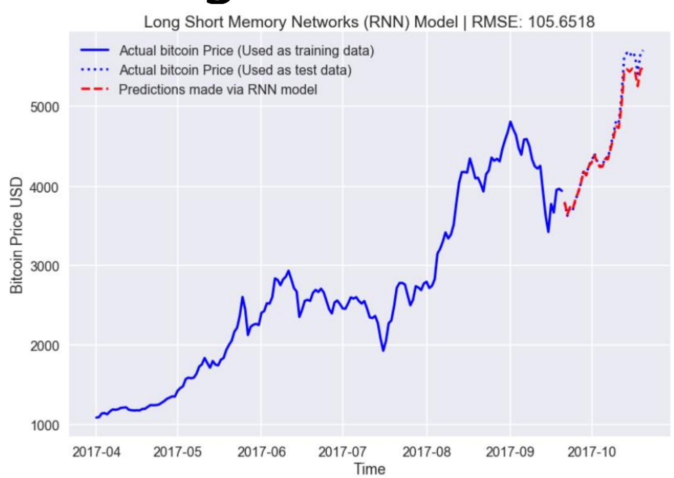


Fig. 7. Cross Validation on a rolling basis [10]

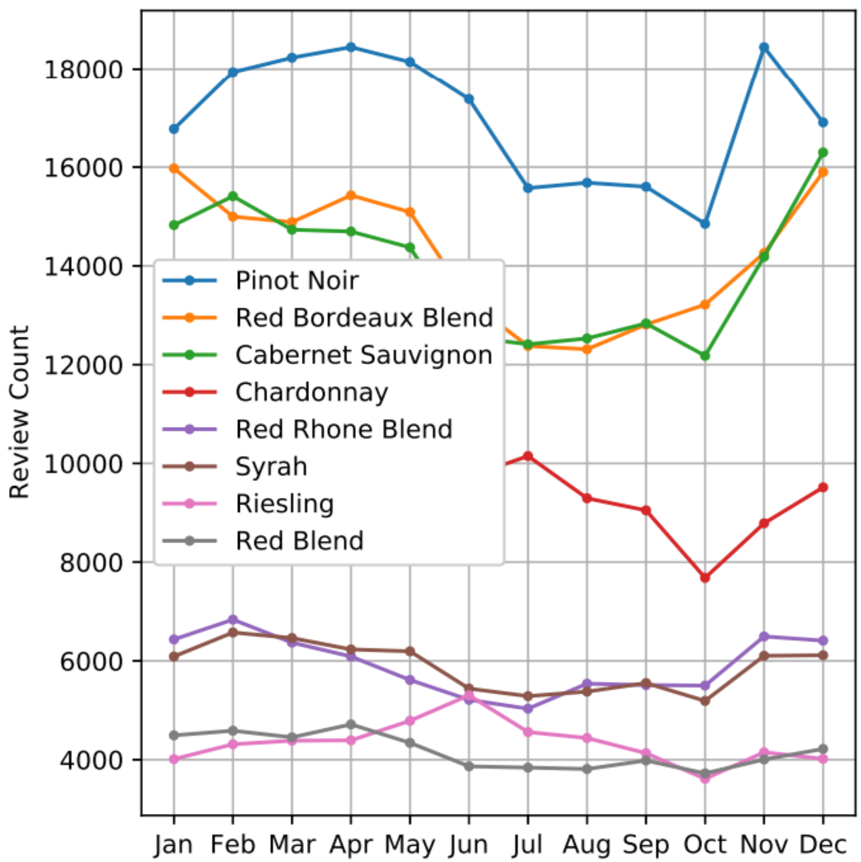
- Does historical Bitcoin data contain enough information to predict its future value (“autoregression”-like task)



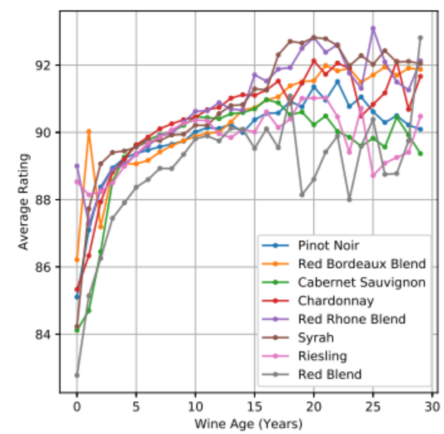
Evaluation Metric	Trained Time Series Models			
	Baseline	***ARIMA***	***Linear Regression***	***LSTM***
RSS	8,529,112	8,148,537	629,980	334,868
MSE	284,303	271,617	20,999	11,162
RMSE	533.20	521.16	144.91	**105.65**

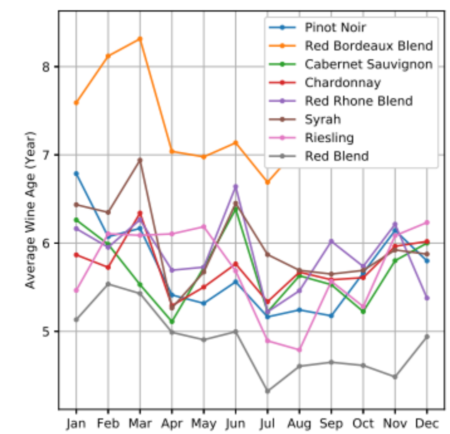
Aman Aggarwal, Gurkanwal Singh Batra

Predicting Wine Popularity Using Temporal Features



- Wine demand appears to exhibit seasonal variability. Can this be predicted?





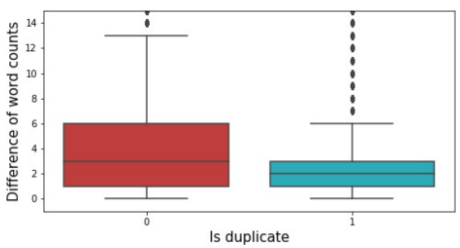
consumption of “high quality” wine is seasonal

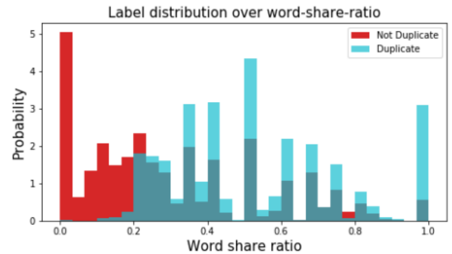
prediction	accuracy
random selection	0.25
pick most popular	0.714
k-nearest neighbor	0.786

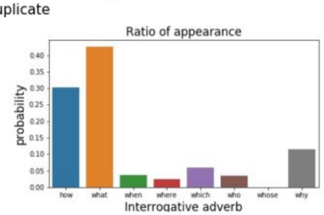
Canruo Ying

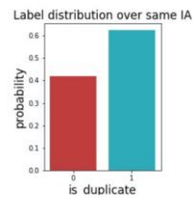
Duplicate Question Detection on Quora

Question1	Question2	label
What can make Physics easy to learn?	How can you make physics easy to learn?	1
What's causing someone to be jealous?	What can I do to avoid being jealous of someone?	0



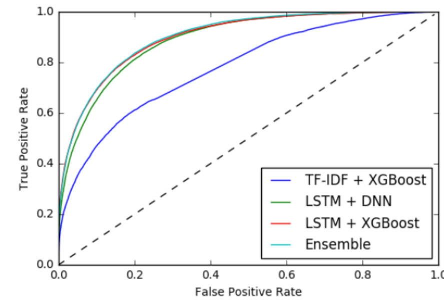






Type	Model	Accuracy
Cosine	Cosine TF-IDF	0.6400
	Cosine topic vector	0.5926
Traditional	LR	0.6405
	SVM	0.6887
	Decision Tree	0.6828
	KNN	0.6769
Ensemble	RF	0.7032
	GBDT	0.7015
	Adaboost	0.6861
Deep model	Siamese LSTM	**0.7754**

Yi Luo,
Jingtao Song,
Haoting Chen



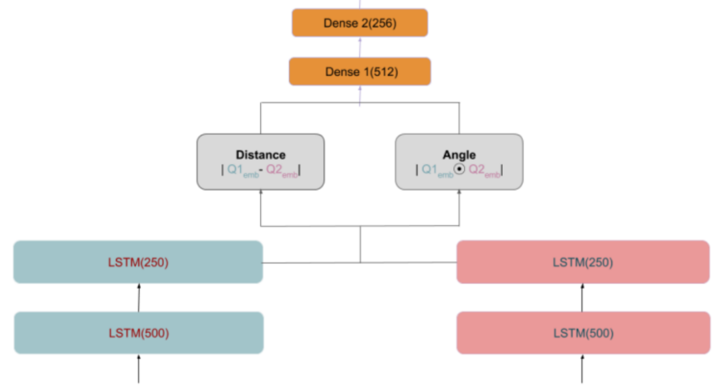


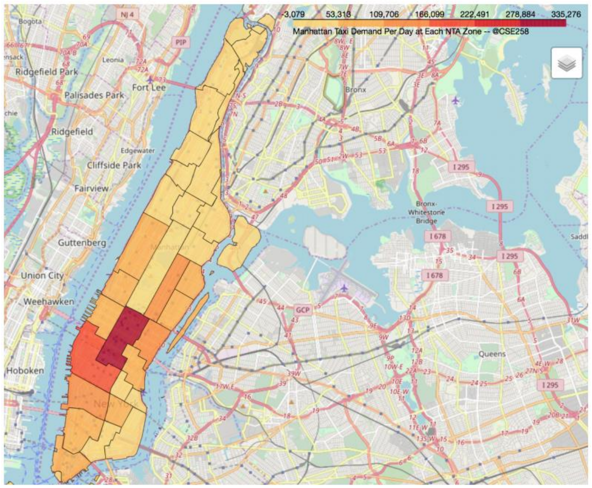
Figure 5: LSTM-based feature extractor followed by handcrafted feature extraction

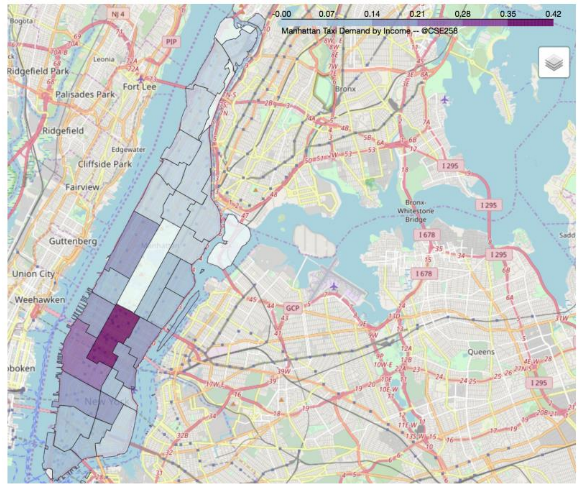
Table 2: Comparative evaluation of all models

Model	Log-Loss	Accuracy(%)	auc	AP
TF-IDF + Cosine Distance	NA	62.9	NA	NA
TF-IDF + XGBoost	0.48	73.66	0.78	0.69
LSTM + DNN	0.39	83.6	0.891	0.83
LSTM + XGBoost	0.38	84.15	0.901	0.851
LSTM + Handcrafted features	0.46	79	0.84	0.82
Ensemble	0.37	84.73	0.903	0.852

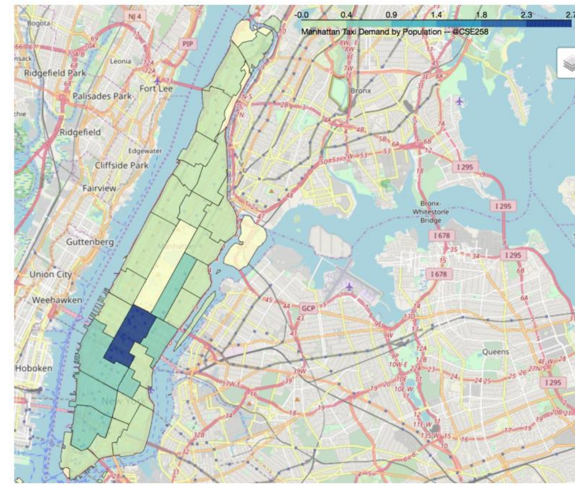
Vaibhav Gandhi, Akshaya Purohit, Aditya Verma

NYC Taxi Demand Prediction

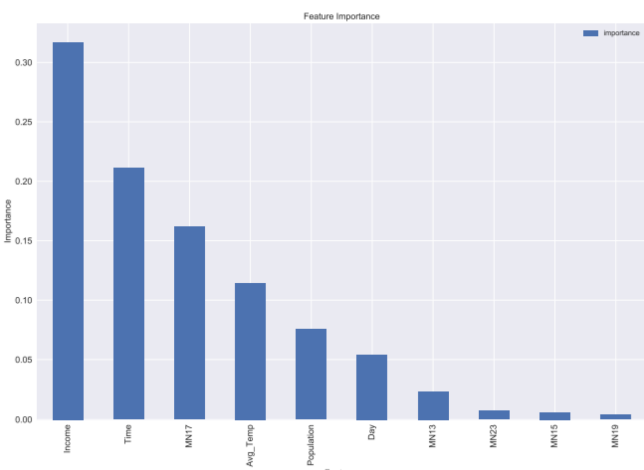




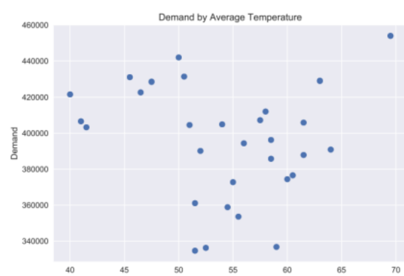
income



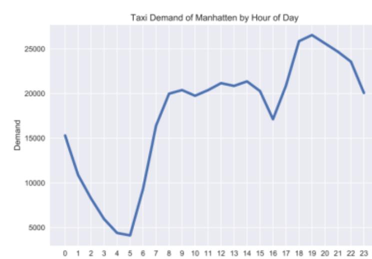
population



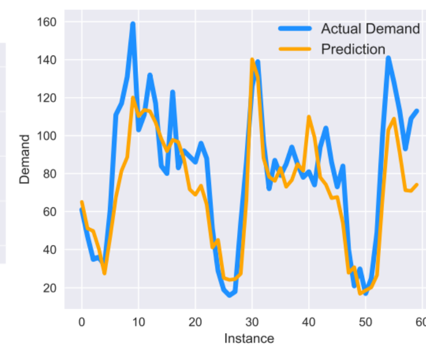
feature importance
(gradient boosted decision tree)



temperature



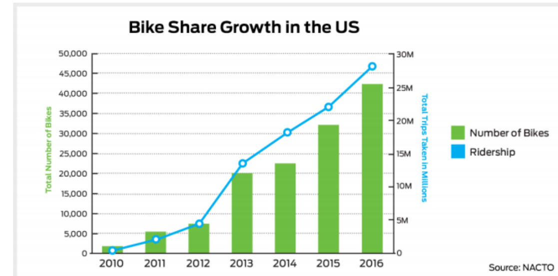
hour



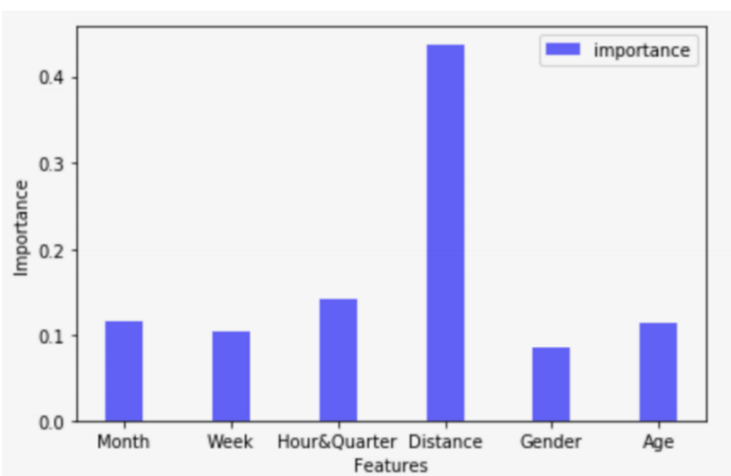
Siyu Jiang, Simran Kapur, Siddharth Dinesh

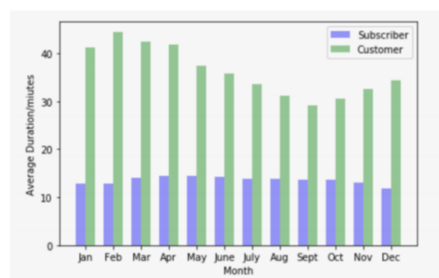
NYC Bike Trip Duration Prediction

Variate	Format
Trip Duration	in seconds format
Start Time and Date	Timestamp
Stop Time and Date	Timestamp
Start Station Name	String
End Station Name	String
Station ID	Number
Station Lat/Long	Number
Bike ID	Number
User Type	Customer or Subscriber
Gender	Number
Year of Birth	Number

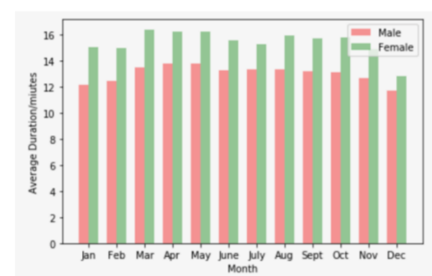


Model	FVU
Baseline	1.000006
Linear Regression	0.211735
Ridge Regression	0.211591
Random Forest Regressor	0.205021
XGBoost Regressor	0.195970
Ensemble of Random Forest and XGBoost	0.200575





subscriber vs. customer



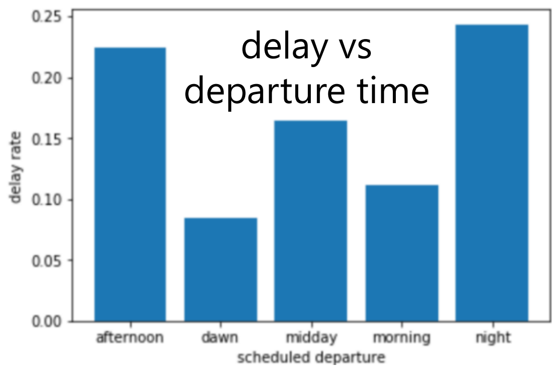
duration vs. gender

Zhuo Cheng, Tianran Zhang, Jiamin He

Airline Delay Prediction

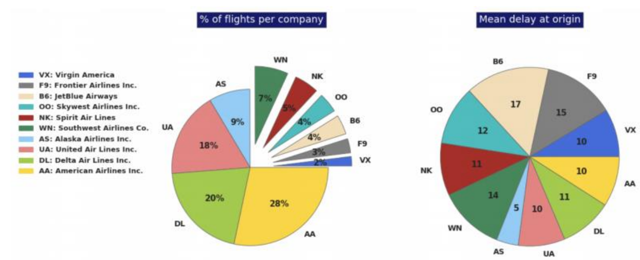
delay vs origin

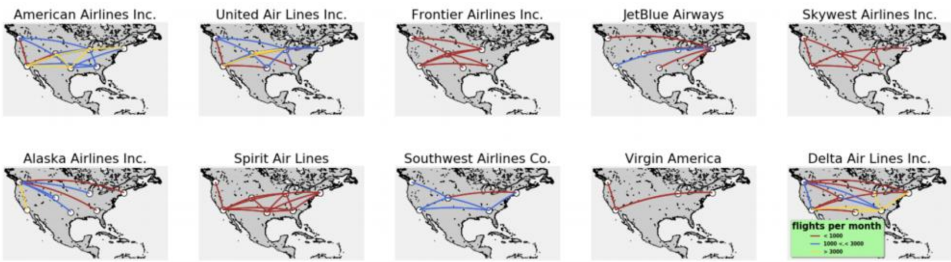
delay vs route



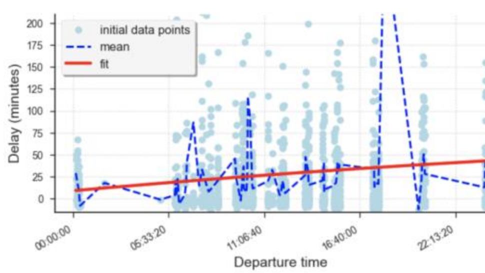
Methods	AUC scores	Precision	Recall	F_1 score	Accuracy
Baseline	0	0	0	0	0.798
Naive Bayes	0.6294	0.3049	0.4044	0.3467	0.6920
Logistic Regression	0.6492	0.3478	0.34	0.3367	0.7345
Random Forest	0.6129	0.2441	0.0074	0.0140	0.7975
Neural Network	0.6404	0.5218	0.0677	0.1150	0.7946

Ran Wang
Qianlong Qu
Yuan Qi
Zijia Chen





Feature Name	Encoding	Dimension
airline	one-hot	10
scheduled_departure	one-hot	24
month	one-hot	12
day_of_month	one-hot	31
day_of_week	one-hot	7
origin_airport	one-hot	7
destination_airport	one-hot	7
distance	float	1
wind_speed	float	1
visibility_in_miles	float	1
sky_coverage	one-hot	5



KNN, SVM, Softmax regression

Qian Zhang
Simeng Zhu
Feng Jiang
He Qin

Fill out those evaluations!

- Please evaluate the course on http://cape.ucsd.edu/students !

Thanks!