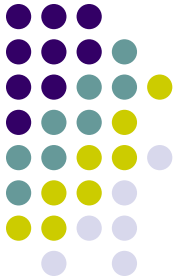
Social Media Analytics

Community Detection
Bi-partite Networks
Cliques and cores

MSITM, 7th November, 2022

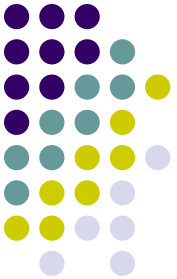
Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor of Business
Distinguished Fellow, INFORMS Information Systems Society
University of Texas Distinguished Teaching Professor
Associate Director, Center for Research in e-Commerce
McCombs School of Business, University of Texas at Austin
Email: aniteshb@gmail.com

Community Detection: Why Bother?

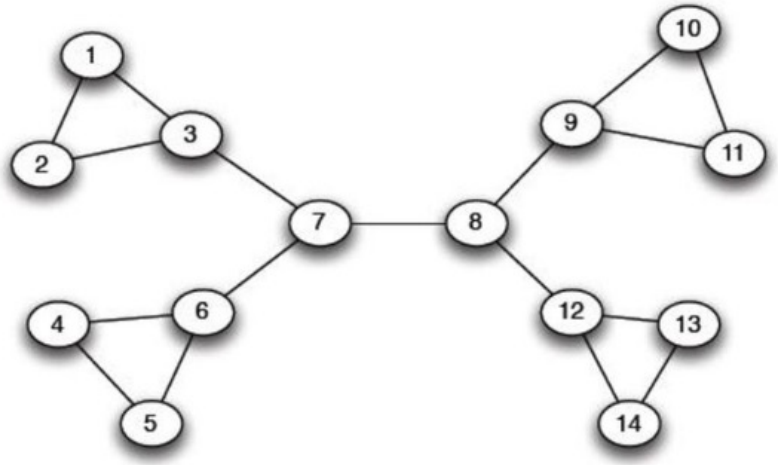


- Detecting networks of fraudulent/rogue websites
  - Many use JavaScript redirects to link to each other to avoid detection through scraping
- Estimating unknown features of users in social networks
- Clustering similar users together
  - Enhance meaningful communication
- Can be a network of products
  - E.g., to show the effect of recommender systems on competition
  - Show that a "community" has products from very different parts of the demand curve
  - https://joshbarua2002.medium.com/who-is-your-competitor-in-the-era-of-the-long-tail-d0ac24fedde8

How to Detect Communities Within Networks



- Common for uni-partite (1-mode) networks
- Girvan-Newman algorithm (divisive algorithm)
- Calculate betweenness centrality of "links"
- How?



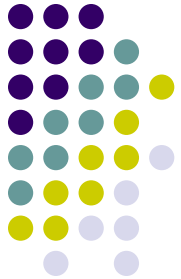
Betweenness(7, 8)= 7x7 = 49

Betweenness(1, 3) = 1X12=12

Betweenness(3, 7) = Betweenness(6, 7) =
Betweenness(8, 9) = Betweenness(8, 12)= 3X11=33

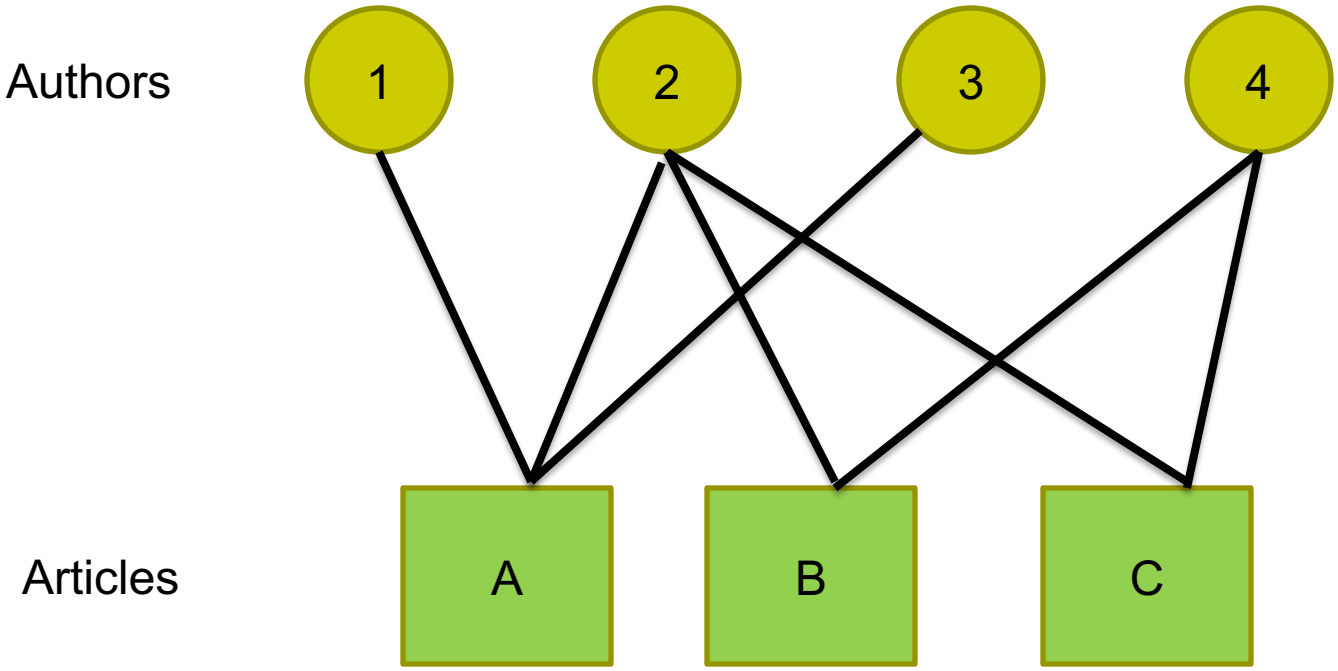
- (i) Cut the link with highest betweenness centrality
- (ii) Recalculate betweenness for all remaining links
- (iii) Cut the link with highest betweenness
- (iv) Repeat (ii) and (iii) until the network disintegrates into disjoint parts
- Excellent article: https://www.analyticsvidhya.com/blog/2020/04/community-detection-graphs-networks/

Types of Networks



- Uni-partite, bi-partite, tri-partite (or multi-partite) networks
- Also called 1-mode, 2-mode, etc.
- Uni-partite: Only one type of nodes (e.g., people)
- Bi-partite: E.g., authors & articles, actors & movies, FB users and their group memberships, etc.

Bi-partite Networks: An Example



No connections between nodes of the same type

Can we reduce this network to 1-mode? How?

2-mode to 1-mode Networks

- 2-mode: Congress(wo)man & age
- How to reduce to 1-mode (person-to-person)?

	Coble	Franks	Goodlatt	Hartzler	McGove	Nadler	Pingree	Polis	Roby	Waters
Coble	0	26	21	29	28	16	24	44	45	7
Franks	26	0	5	3	2	10	2	18	19	19
Goodlatte	21	5	0	8	7	5	3	23	24	14
Hartzler	29	3	8	0	1	13	5	15	16	22
McGovern	28	2	7	1	0	12	4	16	17	21
Nadler	16	10	5	13	12	0	8	28	29	9
Pingree	24	2	3	5	4	8	0	20	21	17
Polis	44	18	23	15	16	28	20	0	1	37
Roby	45	19	24	16	17	29	21	1	0	38
Waters	7	19	14	22	21	9	17	37	38	0

Source: http://www.umasocialmedia.com/socialnetworks/networks-lecture-13-qap-correlation/

Gender & Committees

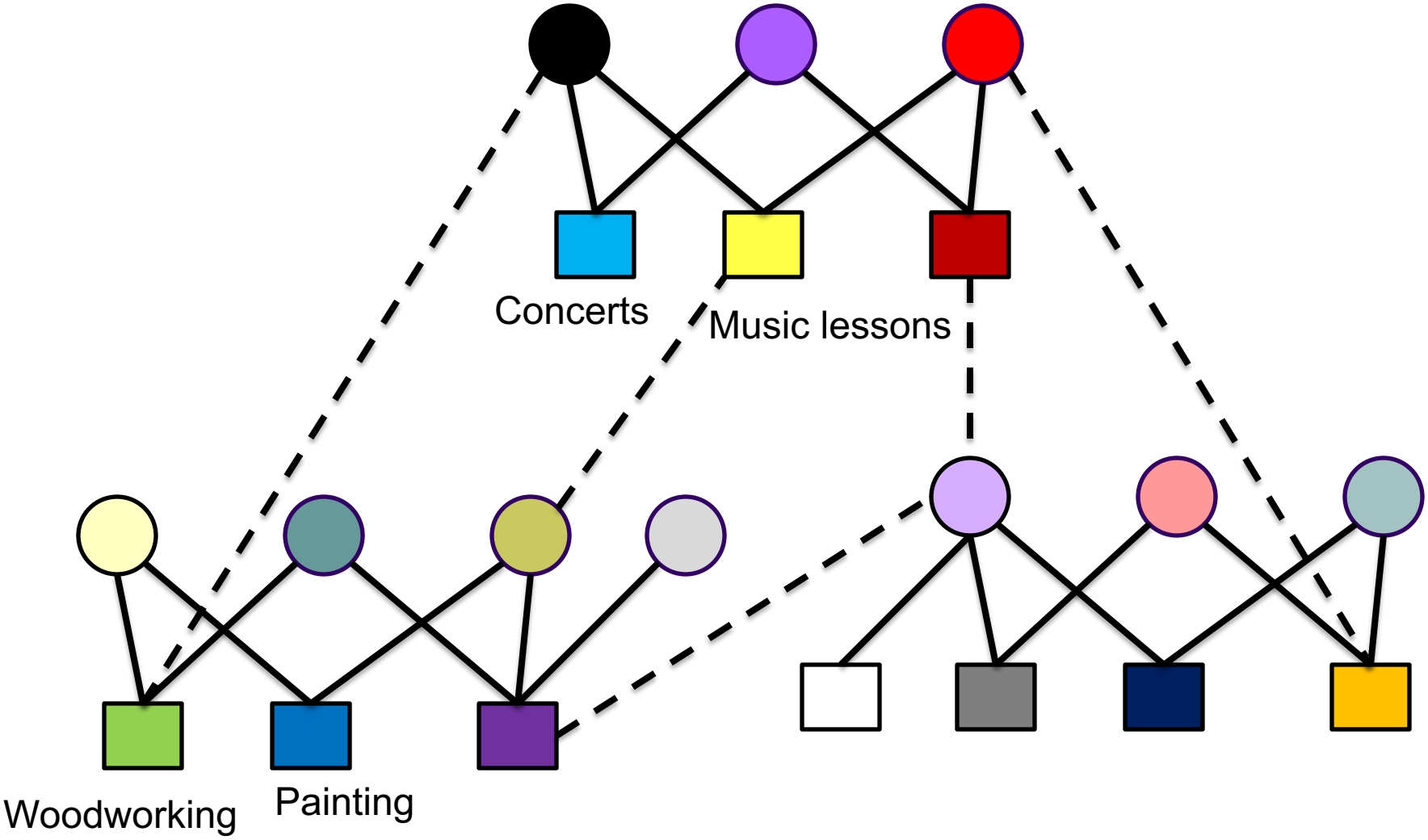
	Coble	Franks	Goodlatt	Hartzler	McGove	Nadler	Pingree	Polis	Roby	Waters
Coble	1	1	1	0	1	1	0	1	0	0
Franks	1	1	1	0	1	1	0	1	0	0
Goodlatte	1	1	1	0	1	1	0	1	0	0
Hartzler	0	0	0	1	0	0	1	0	1	1
McGovern	1	1	1	0	1	1	0	1	0	0
Nadler	1	1	1	0	1	1	0	1	0	0
Pingree	0	0	0	1	0	0	1	0	1	1
Polis	1	1	1	0	1	1	0	1	0	0
Roby	0	0	0	1	0	0	1	0	1	1
Waters	0	0	0	1	0	0	1	0	1	1

	Coble	Franks	Goodlatt	Hartzler	McGove	Nadler	Pingree	Polis	Roby	Waters
Coble	1	1	1	0	0	1	0	1	0	1
Franks	1	2	2	1	0	1	1	1	1	1
Goodlatte	1	2	2	1	0	1	1	1	1	1
Hartzler	0	1	1	2	1	0	2	0	2	0
McGovern	0	0	0	1	2	0	1	1	1	0
Nadler	1	1	1	0	0	1	0	1	0	1
Pingree	0	1	1	2	1	0	2	0	2	0
Polis	1	1	1	0	1	1	0	2	0	1
Roby	0	1	1	2	1	0	2	0	2	0
Waters	1	1	1	0	0	1	0	1	0	1

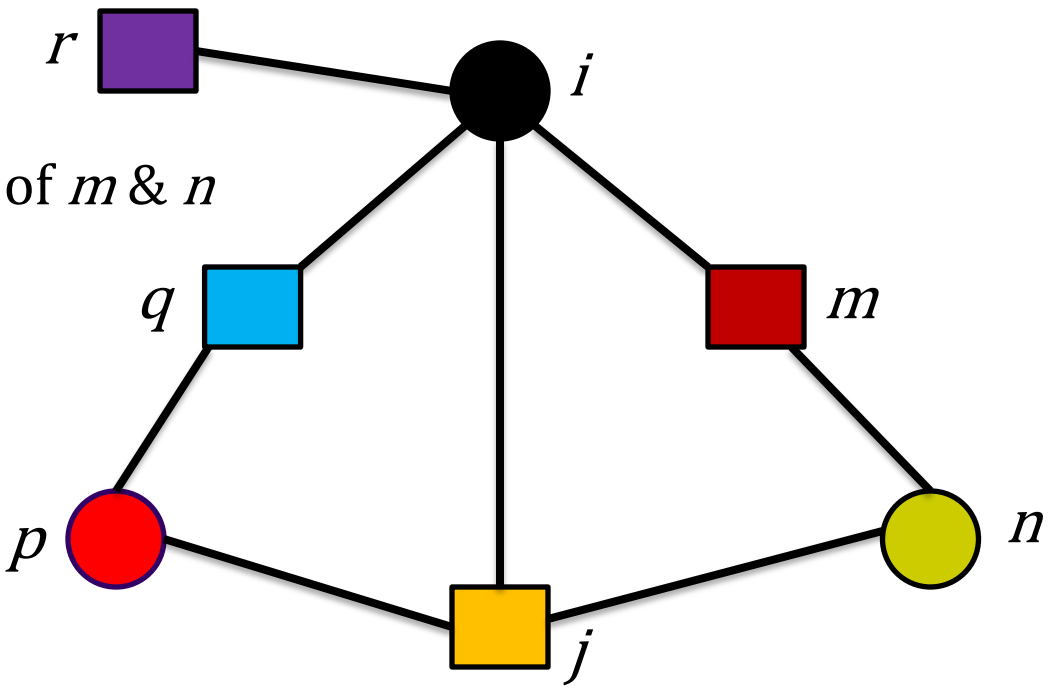
Source: http://www.umasocialmedia.com/socialnetworks/networks-lecture-13-qap-correlation/

Communities in Bi-partite Networks

- Densely linked parts of a bi-partite network constitute communities
- E.g., people's memberships in a set of activities



Detecting Communities in Bi-partite Networks



k_m & k_n degrees of m & n

For a pair of nodes, *i* and *j*, let *m* and *n* be neighbors of *i* and *j* respectively

q_{ijmn} = 1 if *m* and *n* are connected, 0 otherwise. θ_{ijmn} has the opposite definition.

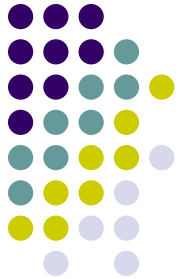
Edge clustering coefficient $C(i,j) =$

squares that currently include *i-j* / possible # squares that include *i-j*

$$\frac{\sum_{m=1}^{k_i}\sum_{n=1}^{k_j} q_{ijmn}}{\sum_{m=1}^{k_i}\sum_{n=1}^{k_j} \theta_{ijmn} + \sum_{m=1}^{k_i}(k_m - 1) + \sum_{n=1}^{k_j}(k_n - 1) - \sum_{m=1}^{k_i}\sum_{n=1}^{k_j} q_{ijmn}}$$

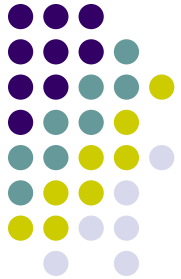
Source: P. Zhang et al. 2013

Detecting Communities in Bi-Partite Networks



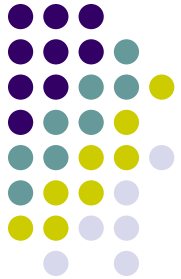
- Start dropping links with smallest clustering coefficient
- Network will start splitting up
- But much easier to first divide into two 1-mode networks
- The apply G-N or other algorithms to detect communities within 1-mode networks
- We do lose some information
- Example: https://medium.com/@adibarua2002/creating-smarter-online-communities-with-nlp-and-network-analytics-147810d3cee5?source=friends_link&sk=d7f5809dfa0e7cc774448615e9287de5

Sample Project on Bi-Partite Networks



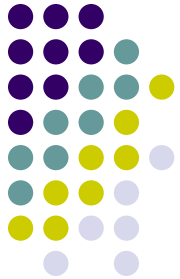
- Readers and books
- How to create a manageable network
- Compare recommendations based on
  - Book-to-book similarity (uni-partite)
  - Person-to-person and then recommending books that similar readers have read but not the focal reader.
- Does the second method provide more variety?

Identifying Trolls, Spammers and Fake Content Creators



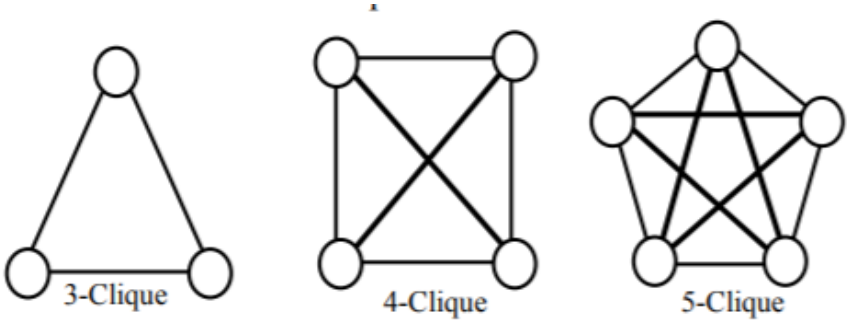
- Trolls, spammers, bots
- Creators of fake content
- Is there anything common across them?
- Was not too difficult to detect fake content (e.g., reviews) from content in the past
- Why is it difficult to detect now?
- Network analytics may help

They Come in Many Flavors



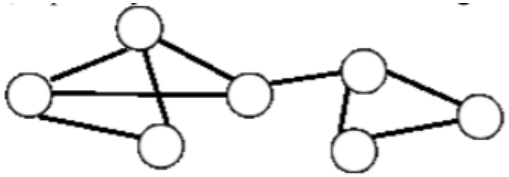
- ~ 19 million bot accounts tweeted in support of either Trump or Clinton in 2016
- > 1,000 paid Russian trolls spread “fake news” on Hillary Clinton
- CNN mobile app received 100s of thousands of 1-star reviews after the network’s treatment of a certain Reddit user
- The Boca Raton Resort hotel got a huge number of negative reviews (1000s) after a Youtube star angry at his treatment rallied his fanbase to retaliate online
- **<u>But may be more connected than regular users!</u>**

Of “Cliques” and “Cores”



- Definition of an n-clique?
- What is the significance of a clique?
- A large network will consist of many cliques
- Often the largest clique is of interest

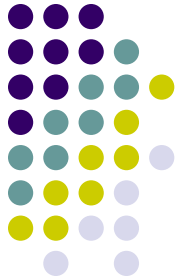
K-core



A 2-core

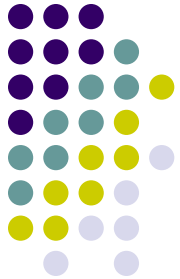
- Definition of a k-core?
- What is the significance?
- Largest k for a network is of special interest

Hypotheses (Can be Tested)



- Spammers have larger k-cores than non-spammers
- Spammers have larger n-cliques than non-spammers
- Spammers have higher network density

Calculating Cliques and k-cores



- Lots of python code available on GitHub and other sites
- https://www.kaggle.com/mayeesha/network-analysis-for-dummies-stackoverflow-data
- https://s3.amazonaws.com/assets.datacamp.com/production/course_3286/slides/ch3_slides.pdf
- https://towardsdatascience.com/intro-to-graphs-in-python-using-networkx-cfc84d1df31f
- Maximum value of k in k-cores within a network
  - https://github.com/chibuta/k-core-subgraph

Data Issues

- If collecting primary data
  - Twitter → Find out important hashtags (e.g., anti- vs. pro-vaxxer, anti-climate change vs. pro)

Stance	Hashtags
Pro-vaccination	*VaccinesSaveLives, VaccinesWork, WorldImmunizationWeek, VaxWithMe, HealthForAll, WiW, ThankYouLaura*
Anti-vaccination	*LearnTheRisk, VaccineInjury, VaccineDeath, VaccineDamage, VaccinesCauseAutism, CDCFraud, CDCWhistleBlower, CDCTruth, WakeUpAmerica, HearUs, HealthFreedom*
Unidentified	*Vaccine, Vaccines, Vaccinate, VaccinateUS*

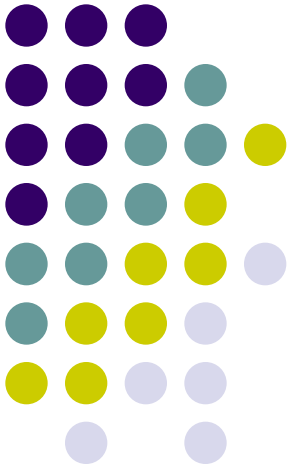
  - Other sources (like discussion forum): Separate source and target during scraping.
    - E.g., creator of an original post in one column
    - Person (or people) commenting in another column
- Many useful sources of archived network data
  - https://github.com/benedekrozemberczki/datasets
  - Other sources on GitHub, some on Kaggle

Unstructured Data Analytics

Homophily vs. Social Influence

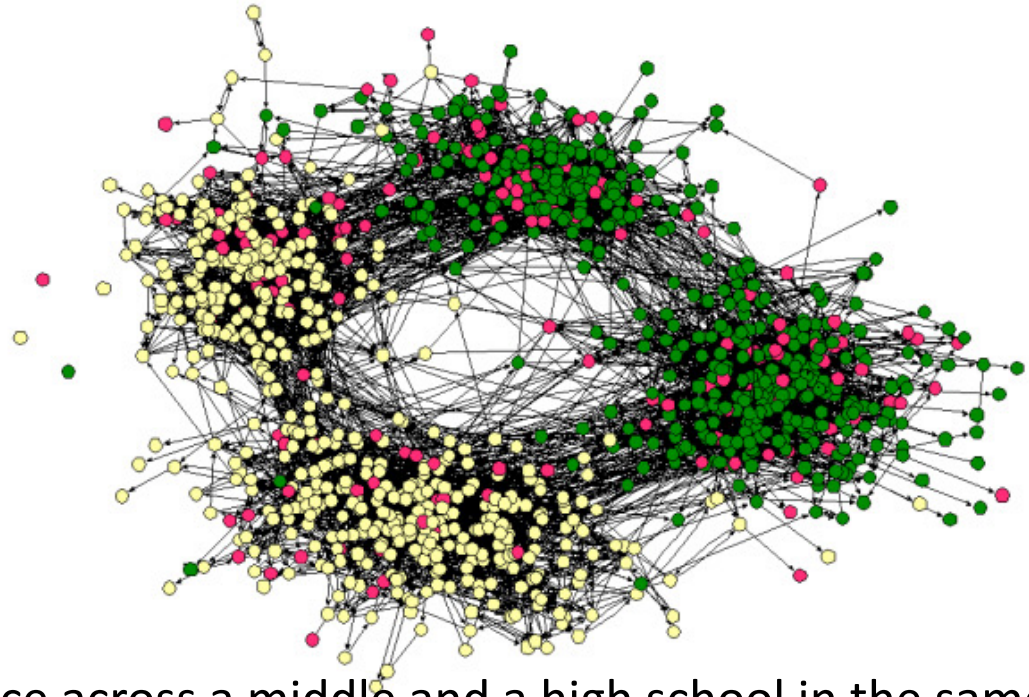
MSITM, F2022, 28th November

Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor of Business
Distinguished Fellow, INFORMS Information Systems Society
Stevens Piper Foundation Professor
University of Texas Distinguished Teaching Professor
Associate Director, Center for Research in e-Commerce
McCombs School of Business, University of Texas at Austin
Email: aniteshb@gmail.com



Homophily (Similarity)

- "Birds of a feather flock together"
  - Your friends/contacts vs. a random sample of people
  - Social networks tend to connect people who are similar to each other



Friendships by race across a middle and a high school in the same school district

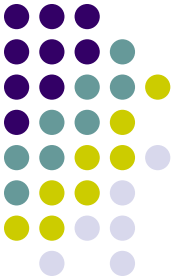
Source: Easley & Kleinberg

Does It Matter?

- Why do websites show ratings by your friends (if you use Facebook to log on)?
- Does homophily play a role? How?
- Pepsi social media chief sums up his strategy in one word: 'Homophily'*

Source: http://www.businessinsider.com/pepsi-social-media-chiefs-strategy-homophily-2012-9#ixzz2sF2Ls7v6

Distinguishing Between Social Influence and Homophily

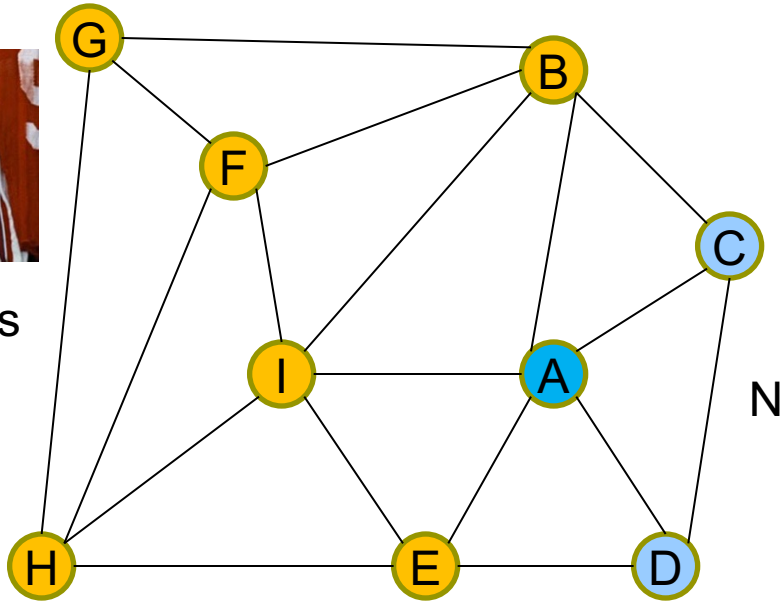


- Can opinions, attitudes & purchases be attributed to social influence?
- Or is it due to homophily?
- E.g., did I buy something because
  - you influenced me?
  - we are just similar?
- What difference would it make to a company's strategy?

Detecting Homophily for Static Attributes

- Have to know what attribute(s) may be relevant
- E.g., gender, interest, educational background, etc.

Jocks



Nerds

- Does this network exhibit homophily?
- What measure can we use?

Source: Easley & Kleinberg

A Little Theoretical Detour

A network with a set of nodes (V) & randomly assigned edges (E^r):

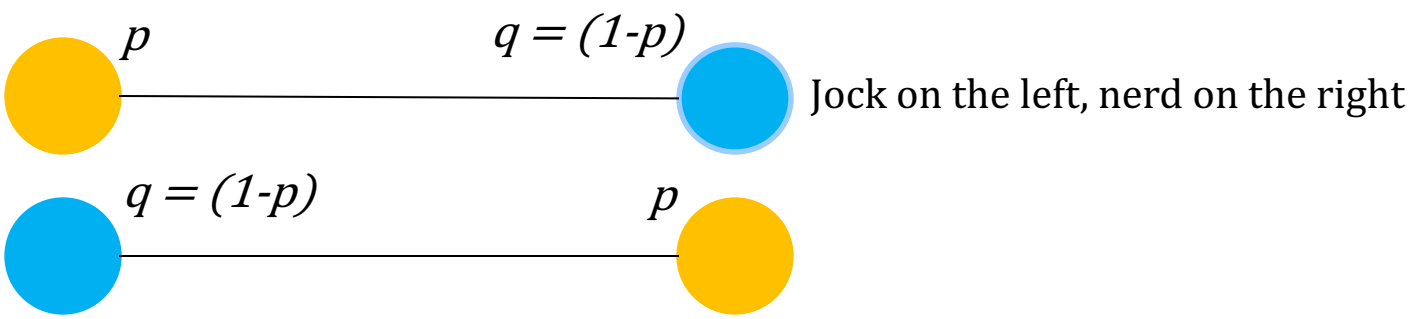
$$R = (V, E^r)$$

Each node is assigned an attribute: say, type = jock with probability p, and type = nerd with probability $q = 1\text{-}p$

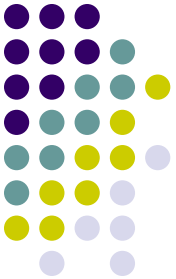
Consider any edge $(i,j) \in E^r$ of this random network R.

Let the random variable $X_{ij} = 1$ if it is a "cross-edge", and $X_{ij} = 0$ otherwise. Then X_{ij} is a Bernoulli random variable such that

$P(X_{ij} = 1) = 2pq$



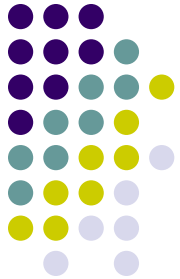
Dynamic Attributes: How Can We Distinguish Between Homophily & Influence?



- Need multiple snapshots in time
- Homophily: Due to **similar** attributes in time *t*, some people may choose to become friends in *t*+1
  - E.g., high achievers in a class may form links
- But some people may become friends in *t*+1 even though their attributes were different in *t*
- Check which effect is stronger

Source: LaFond & Neville

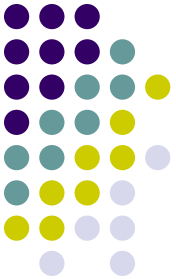
Test for Homophily



- Homophily exists if
- *p*(Becoming friends in *t*+1 where attributes were same in *t*) > *p*(Becoming friends in *t*+1 where attributes were different in *t*)
- *p*(Dissolving friendships in *t*+1 where attributes were same in t) < *p*(Dissolving friendships in *t*+1 where attributes were different in *t*)

Source: LaFond & Neville

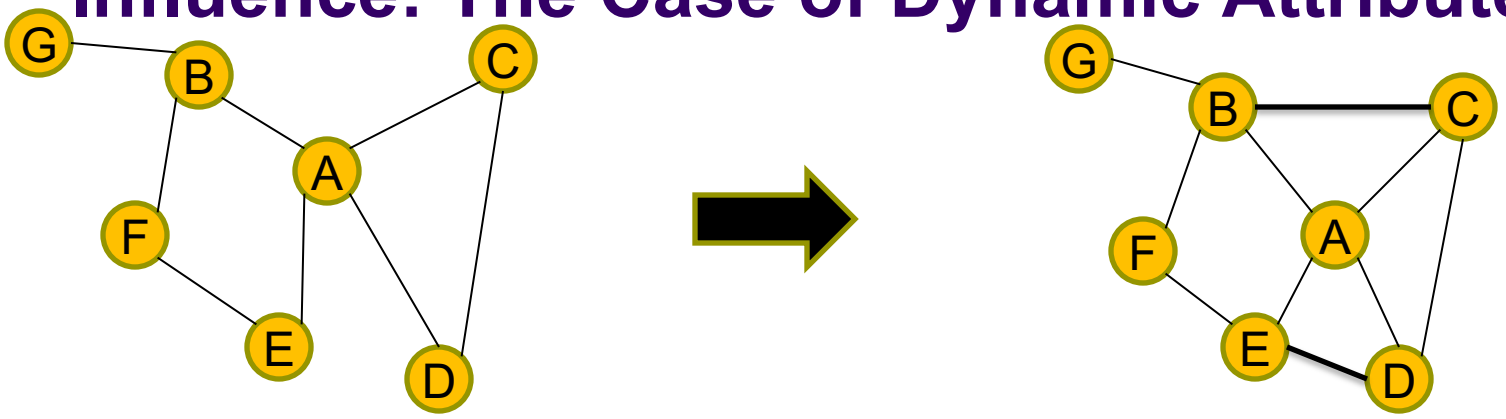
Detecting Social Influence



- Some **friends** at *t* with different attributes may become similar in *t*+1 (due to social influence)
  - E.g., some buy a product their friends have
  - Some change their beliefs & attitudes
- But people who are not friends and have different attributes at *t* can also become similar at *t*+1 due to "other" factors
- Which effect is stronger?
  - I.e., is *p*(Attributes becoming same in *t*+1 where the individuals were friends in *t*) > *p*(Attributes becoming same in *t*+1 where the individuals were not friends in *t*)?
  - Is *p*(Attributes becoming different in *t*+1 where the individuals were friends in *t*) < *p*(Attributes becoming different in *t*+1 where the individuals were not friends in *t*)?

Source: LaFond & Neville

Distinguishing Between Homophily & Social Influence: The Case of Dynamic Attributes



Time: *t*

t+1

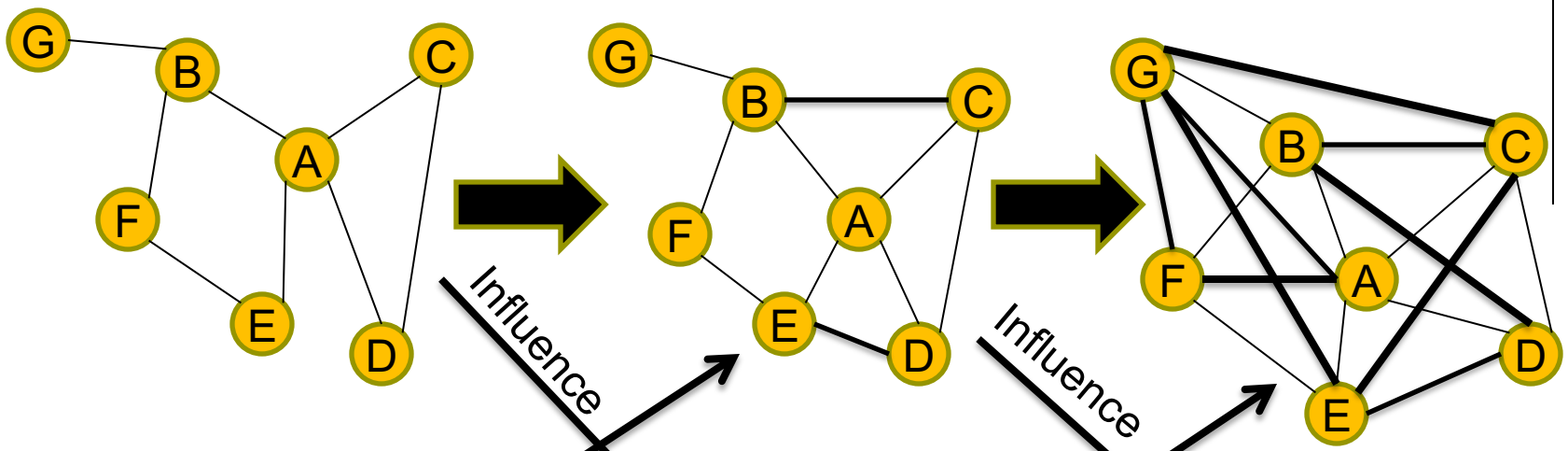
	Subscription
A	Yes
B	No
C	No
D	Yes
E	No
F	Yes
G	Yes

Attribute (e.g., subscription to a music service at time *t*

A	Yes
B	Yes
C	Yes
D	Yes
E	No
F	Yes
G	Yes

Attribute at time t+1

Evidence of Homophily Versus Social Influence



Time: *t*

t+1

t+2

Influence

Homophily

Influence

Homophily

A	Yes
B	No
C	No
D	Yes
E	No
F	Yes
G	Yes

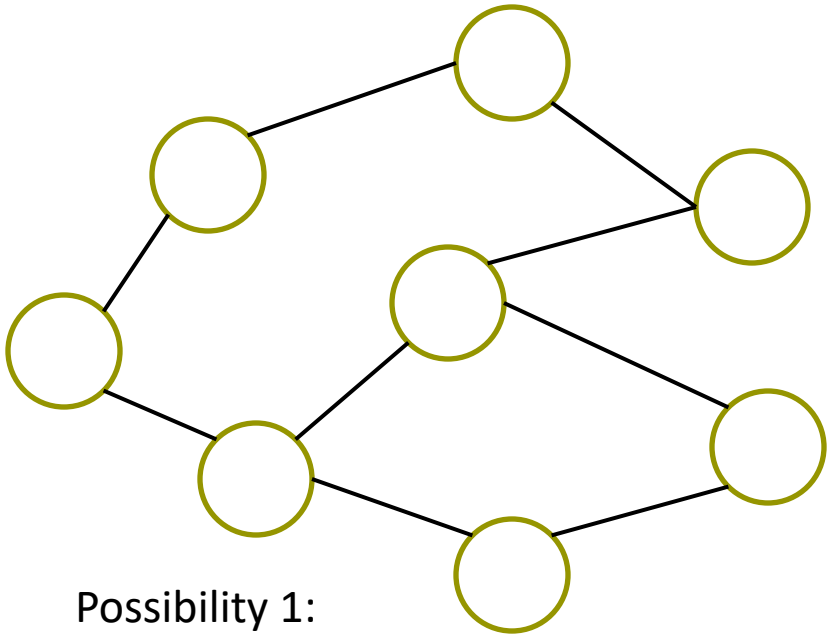
Attribute at time *t*

A	Yes
B	Yes
C	Yes
D	Yes
E	No
F	Yes
G	Yes

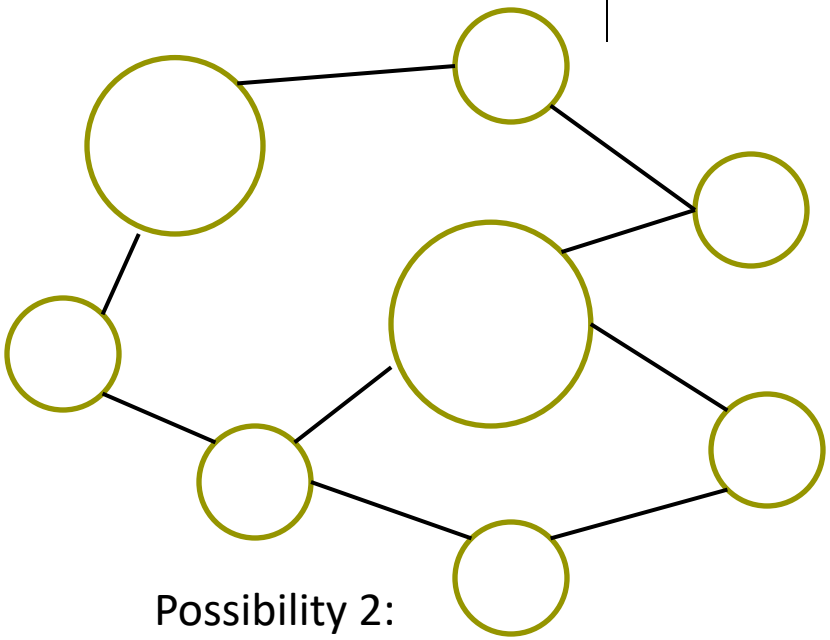
Attribute at time *t*+1

Attribute at time *t*+2

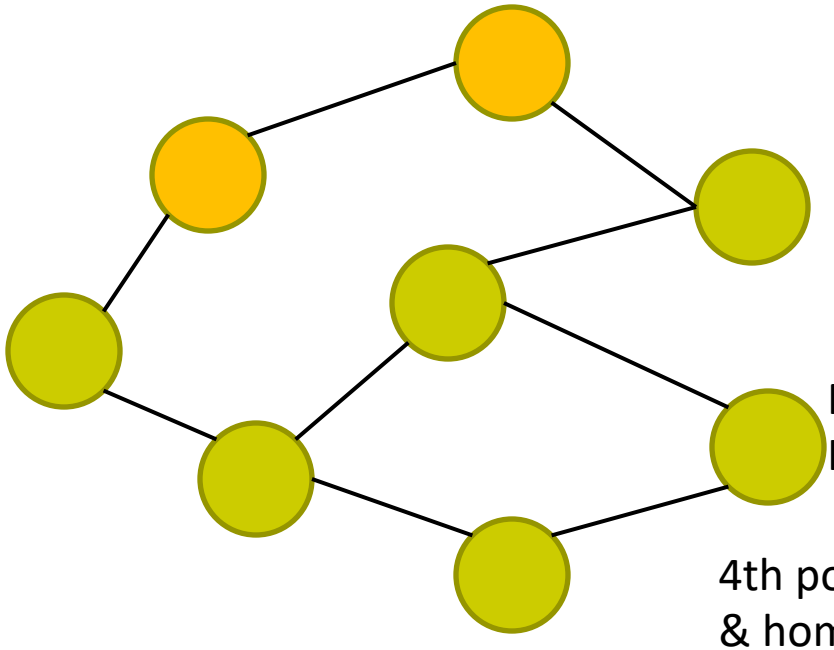
Not all Networks are Created Equal



Possibility 1:
No social influence, no homophily



Possibility 2:
Social influence but no homophily



Possibility 3:
Homophily but no social influence

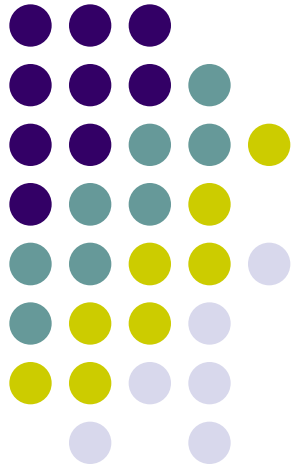
4th possibility (not shown): Both influence & homophily

Unstructured Data Analytics

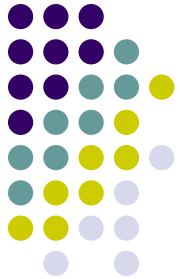
Text Clustering & Topic Modeling

MSITM, F2022, 3rd October

Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor of Business
Distinguished Fellow, INFORMS Information Systems Society
Stevens Piper Foundation Professor
University of Texas Distinguished Teaching Professor
Associate Director, Center for Research in e-Commerce
McCombs School of Business, University of Texas at Austin
Email: aniteshb@gmail.com

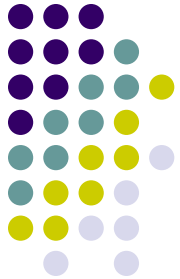


Document Clustering



- Useful in reducing a large number of diverse documents into groups, topics and themes.
- Often uses cluster centroid to assess intra- and inter-cluster similarity
- Euclidean distance or cosine similarity commonly deployed
- Unsupervised method

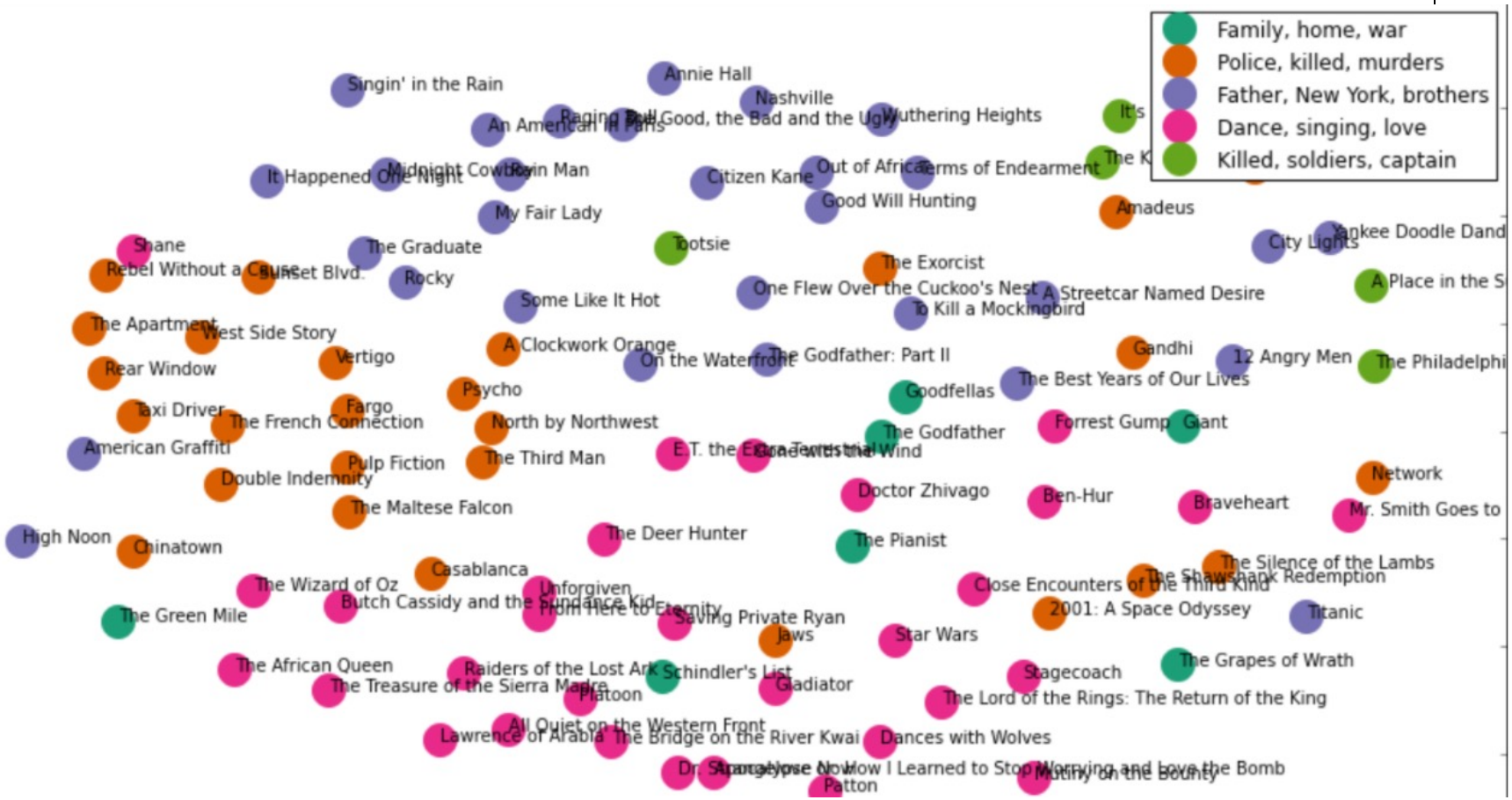
An Example: Clustering Top 100 Movies Based on Synopses



- Tokenize and stem/lemmatize each synopsis
- Transform corpus into vector space using tf-idf
- Calculating Euclidean distance or cosine similarity between each document
- Cluster using k-means clustering
- Use MDS and visualize with distance or similarity (not a part of clustering)

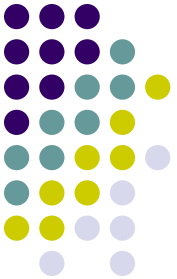
Source: http://brandonrose.org/clustering

Clustering & MDS of IMDb Top 100 Movies



Source: http://harrywang.github.io/document_clustering/

External Evaluation of Clusters (for Classification)



- Relevant when class information is available

n = number of documents

q = number of classes in data (e.g., negative, neutral, positive)

n^i_r is the number of documents of the ^{i}th class assigned to ^{r}th cluster S_r

n_r is the # documents in cluster S_r

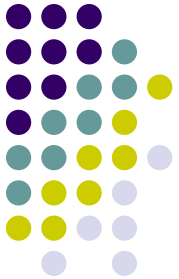
k clusters

Entropy of S_r is $$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}$$

Entropy of clustering is $$entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r)$$

Observations?

An Alternative to Entropy

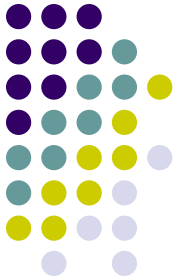


Cluster purity: What % of a cluster belongs to the dominant class?

Purity of a cluster S_r $$Pu(S_r) = \frac{1}{n_r} \max_i n_r^i$$

Purity of a clustering $$purity = \sum_{r=1}^{k} \frac{n_r}{n} Pu(S_r)$$

A Different Measure



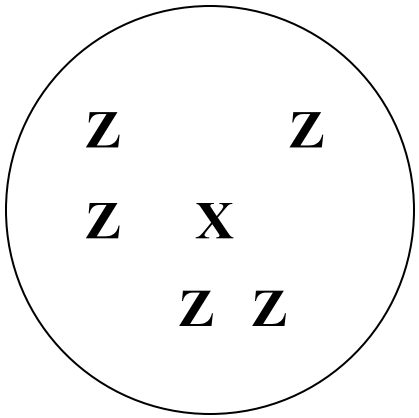
- Consider searching for documents of a particular class. Clustering yields the results.
- TP: Assigns two _____ docs to _____ cluster(s)
- TN: Assigns two _____ docs to _____ cluster(s)
- FP: Assigns two _____ docs to _____ cluster(s)
- FN: Assigns two _____ docs to _____ cluster(s)
- Rand Index (RI) = (TP + TN) / (TP + FP + TN + FN)

Source: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

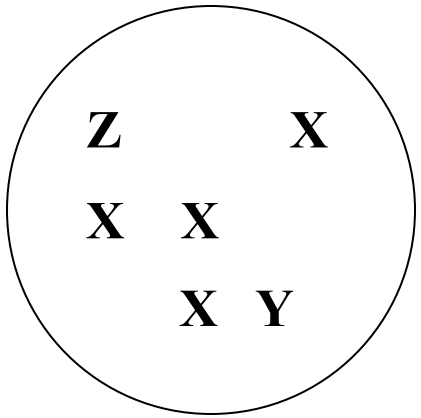
Rand Index (RI) Calculation

Number of document pairs	Assigned to same cluster	Assigned to different clusters
Same class in ground truth	TP	FN
Different classes in ground truth	FP	TN

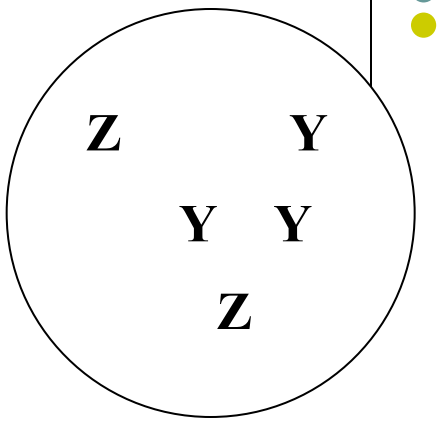
An Example



Cluster I



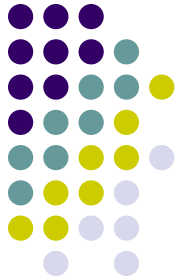
Cluster II



Cluster III

- TP = ?
- FP = ?
- TN = ?
- FN = ?
- RI = ?

1 Document, 1 Topic?



- What are the topics in these documents?
  - "I like to eat broccoli and bananas."
  - "I ate a banana and spinach smoothie for breakfast."
  - "Chinchillas and kittens are cute."
  - "My sister adopted a kitten yesterday."
  - "Look at this cute hamster munching on a piece of broccoli."
- Latent Dirichlet Allocation (LDA): a way of automatically discovering **topics** that these sentences contain
- Topics and their distributions are "latent"

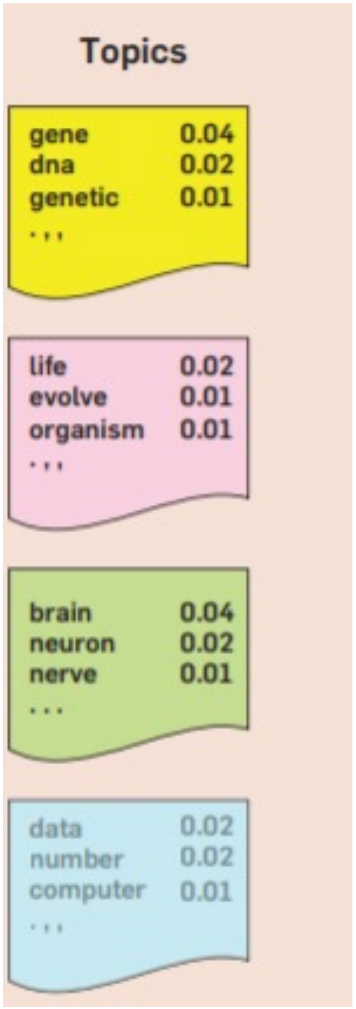
Source: E. Chen

The Key Ideas

- **Topics and words**
  - **Topic A**: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, …
  - **Topic B**: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, …
- **Documents and topics**
- **Documents 1 and 2**: 100% Topic A
- **Documents 3 and 4**: 100% Topic B
- **Document 5**: 60% Topic A, 40% Topic B

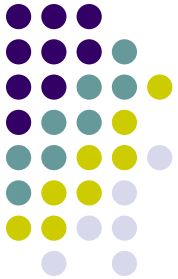
Source: E. Chen

A Simple Example



Source: D. Blei

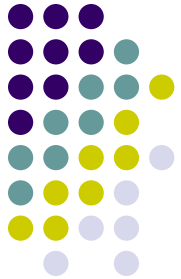
LDA Assumptions



- LDA represents documents as **mixtures of topics**
- Topics contain words with certain probabilities
- Documents were created as follows:
- For each document, the author decided on
  - the number of words *N* (e.g., according to a Poisson distribution)
  - a topic mixture (according to a Dirichlet distribution over a set of *K* topics)
  - E.g., a document may have 1/3 **food** topic and 2/3 **cute animals** topic
- The author generated each word w_{id} in document *d* by:
  - Choosing a topic (according to a multinomial distribution)
  - E.g., **food** topic with 1/3 probability & **cute animals** with 2/3 probability
  - Generate the word itself (according to the topic's multinomial distribution).
  - E.g., for the **food** topic, generate the word "broccoli" with 30% probability, "bananas" with 15% probability, etc.
- Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.

Source: E. Chen

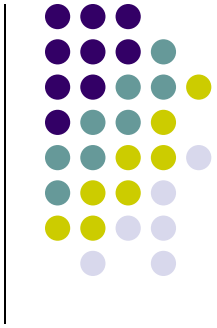
Using Collapsed Gibbs Sampling



- You have a set of documents.
- Choose *K*, the number of topics to discover.
- Go through each document, and use the Dirichlet distribution to assign each word in the document to one of the *k* topics.
- Can also do the initial assignment randomly, but it won't be LDA anymore

Source: E. Chen

Iterations to Reassign Topics to Words



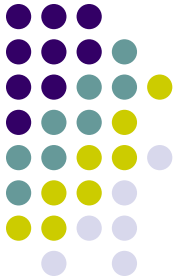
Topic	Words in document X
Food	Fish
Food	Fish
Food	Eat
Food	Eat
Food	Veggies

Topic	Words in document Y
?	Fish
Food	Fish
Food	Milk
Pet	Kitten
Pet	Dog

- For each word *w* in *d* and each topic *t*, calculate
- *p*(*w* | *t*) = proportion of assignments to topic *t* over **all** documents that come from this word *w*.
- *p*(*t* | *d*) = proportion of words in document *d* that are currently assigned to topic *t*
- Reassign *w* in *d* a new topic *t* with probability: *p*(*t* | *d*) * *p*(*w* | *t*)
- This is the probability that topic *t* generated word *w* in document *d*
- Resample the current word's topic with this probability
- Assumption: All topic assignments except for the current word in question are correct

Source: E. Chen

Clustering Vs. Topic Modeling



- A document must belong to one cluster only
- Cluster described by words unique to that cluster

Important words:
politics, election, president

Doc 1
Doc 3

Cluster 1

Important words:
jobs, rust, outsourcing

Doc 2
Doc 4
Doc 6
Doc 7
Doc 10

Cluster 2

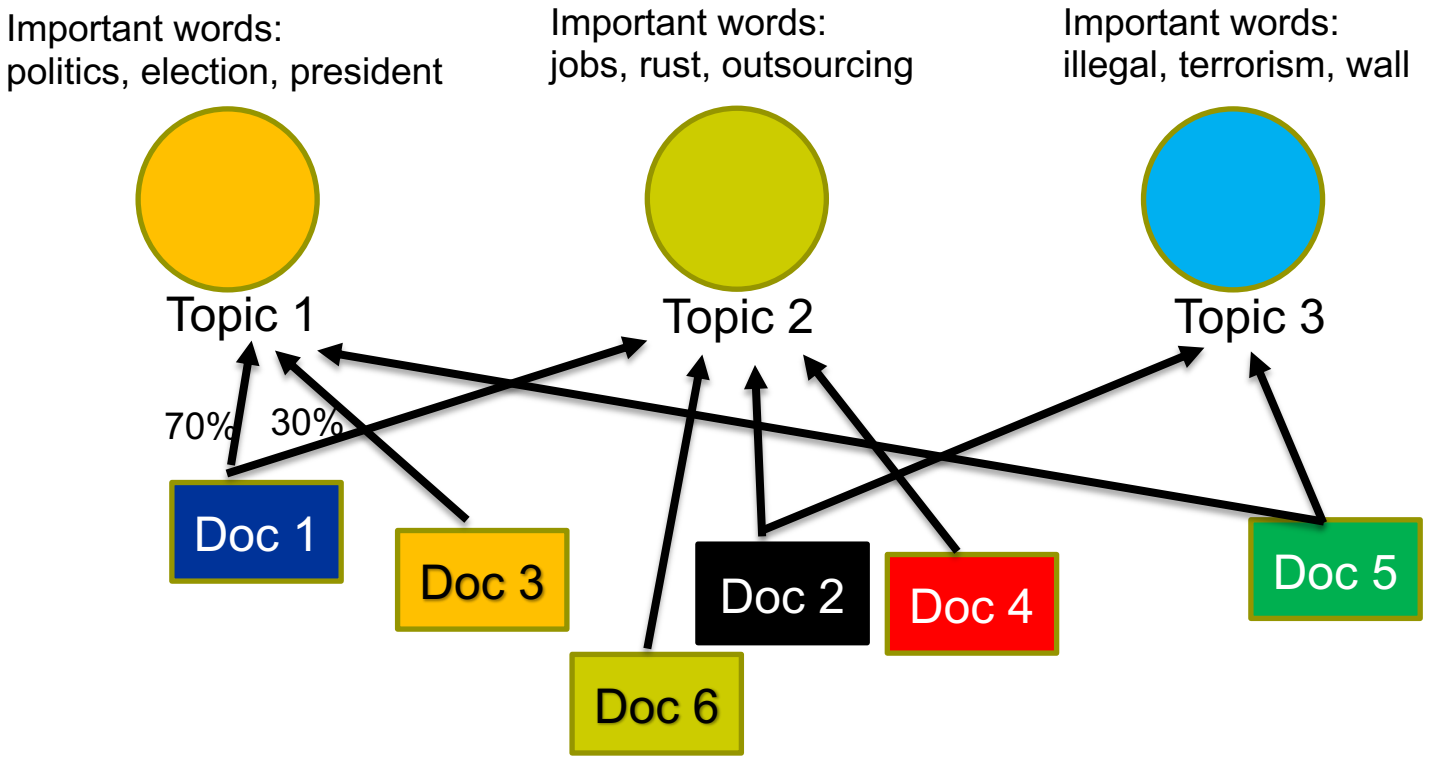
Important words:
illegal, terrorism, wall

Doc 5
Doc 8
Doc 9

Cluster 3

Topic Modeling

- Completely different approach
- May come up with the same keywords as clustering
- But a document can have more than one topic



Unstructured Data Analytics

Image Analytics

MSITM, F2022, 10th October

Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor of Business
Distinguished Fellow, INFORMS Information Systems Society
Stevens Piper Foundation Professor
University of Texas Distinguished Teaching Professor
Associate Director, Center for Research in e-Commerce
McCombs School of Business, University of Texas at Austin
Email: aniteshb@gmail.com

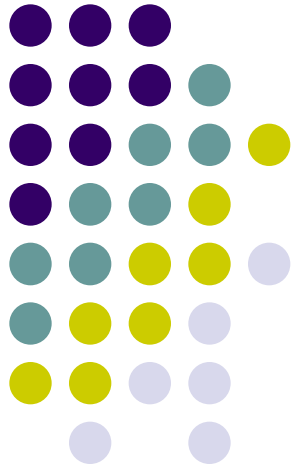
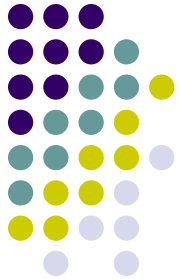


Image Analytics: Why Bother



- Images (and videos) contribute much more to big data than text (or numbers)
- > 4B images posted per day on social media
- > 80% have no text attached
- Medical images in healthcare
- Can provide additional insights

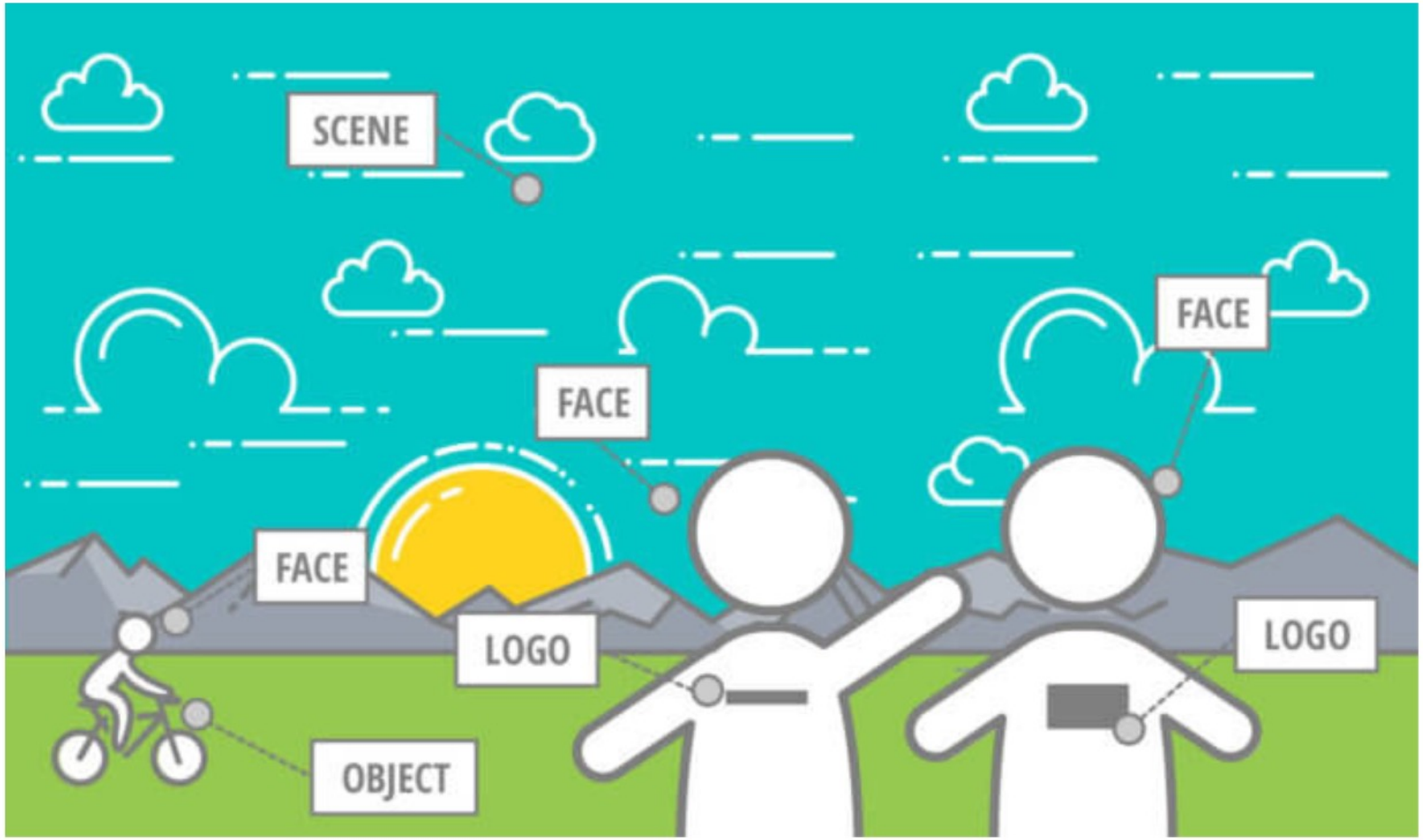
What Does an Image Reveal?

Dubai you will be missed. Thank you for the inspiration and the reset. Back to work now. 💙🌺

Post by Skrillex

Crimson Hexagon

Many Elements



Crimson Hexagon

Moments of “Truth”

Whole > Sum of Parts?

Crimson Hexagon

cigarkirk
@cigarkirk

Anguilla - beach & beer! (@ Blanchards Beach Shack in Meads Bay, Anguilla) swarmapp.com/c/kz1oUtx3QUK

9:53 AM - Dec 15, 2016

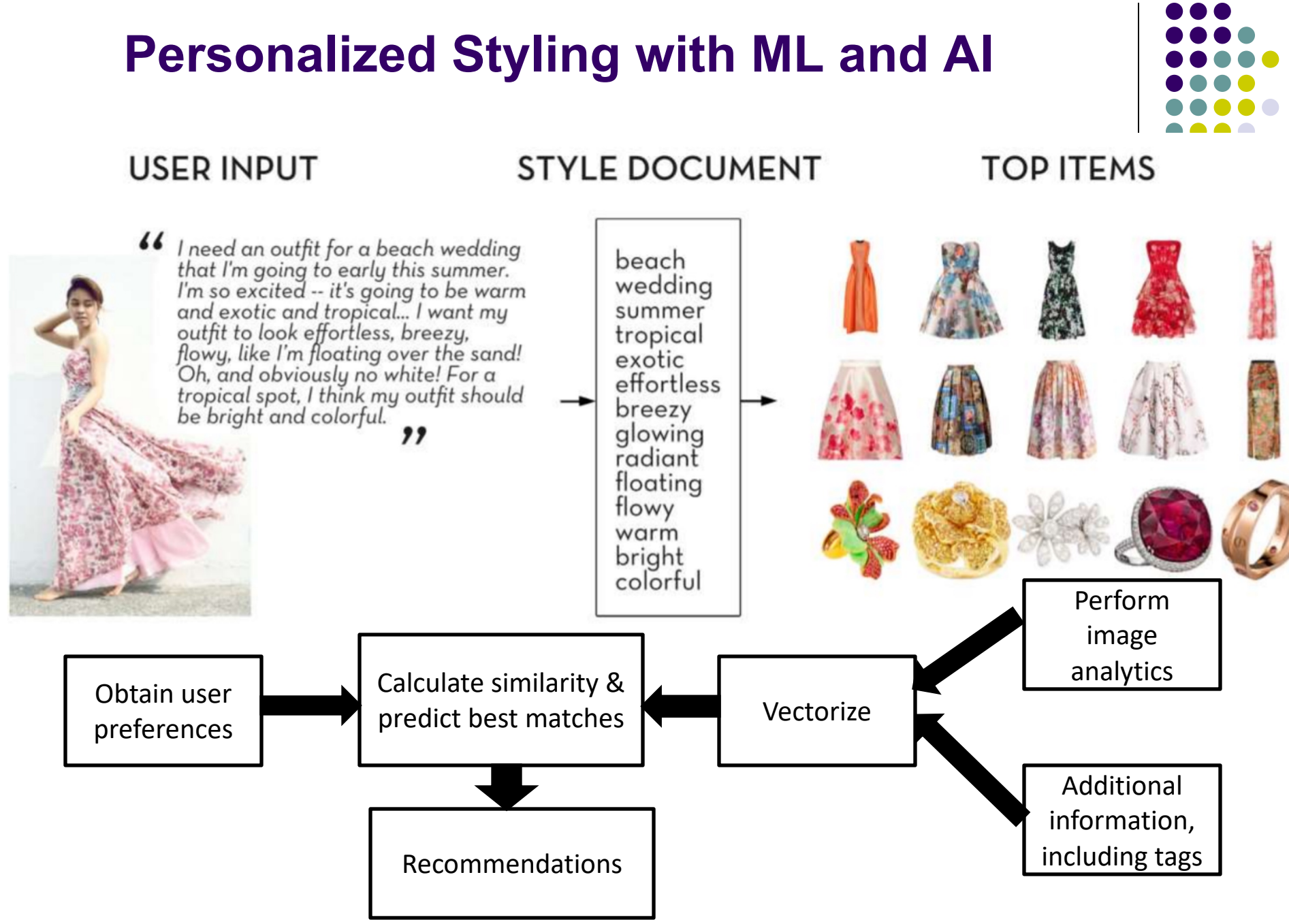
Cookie Takanawa
@JDaddy_3000

Gross flavor. Highly disappointed.

11:42 PM - Jun 2, 2016

Crimson Hexagon

Personalized Styling with ML and AI



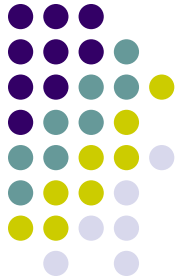
A Quick Overview of Convolutional Neural Networks (CNN): A Computer's View of an Image

25	2	1	44
223	7	6	60
196	8	2	148
249	1	3	40
60	7	1	154
59	1	7	213
214	7	3	163
89	182	219	13
74	146	113	72
89	18	244	85
1	4	8	97
3	4	2	121
2	1	2	131
7	6	8	47
3	5	5	126
7	6	8	121
5	3	1	237

- Pixel representation
- Single "channel"
- "1" for white
- "255" for darkest shade of green

Source: D. Gupta, 2017

How Would a Traditional NN Recognize the Number?



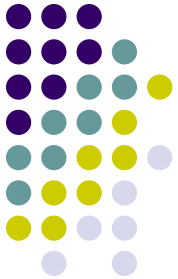
25	2	1	44
223	7	6	60
196	8	2	148
249	1	3	40
60	7	1	154
59	1	7	213
214	7	3	163
89	182	219	13
74	146	113	72
89	18	244	85
1	4	8	97
3	4	2	121
2	1	2	131
7	6	8	47
3	5	5	126
7	6	8	121
5	3	1	237

- "Flatten" the pixel data
- Each pixel an input
- Problems with this approach?

25	223	196	249	60	.....	47	126	121	237

Source: D. Gupta, 2017

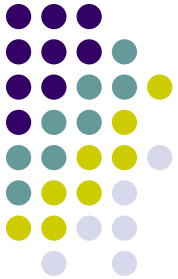
Extracting Features of the Image



25	2	1	44
223	7	6	60
196	8	2	148
249	1	3	40
60	7	1	154
59	1	7	213
214	7	3	163
89	182	219	13
74	146	113	72
89	18	244	85
1	4	8	97
3	4	2	121
2	1	2	131
7	6	8	47
3	5	5	126
7	6	8	121
5	3	1	237

Source: D. Gupta, 2017

Retaining Spatial Information



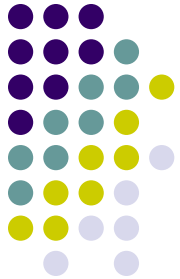
Need to send a 2D (or 3D) arrangement of pixel values to the NN

Take 2 or more pixel values at a time, and suitable weights

What changes do you see on the right?

Source: D. Gupta, 2017

A Possible Solution to the Edge Visibility Problem



Use greater weight on the right
E.g., .1, 5
Combine the two processed images later

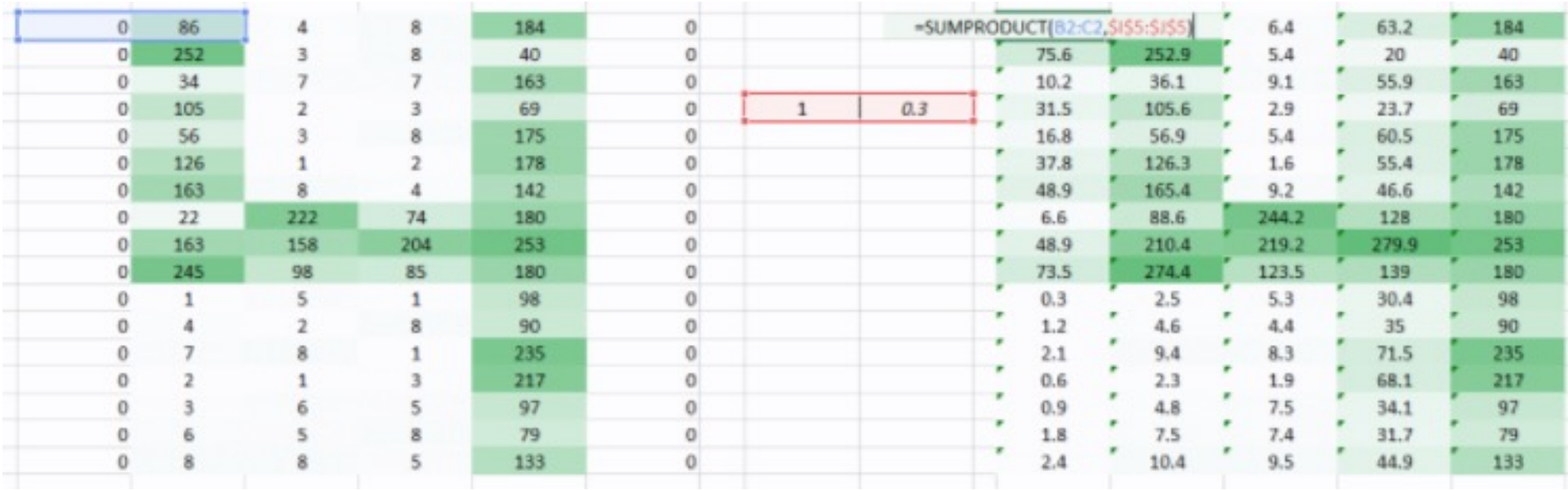
28.6	40.4	920.8
40.2	40.3	200.8
38.4	35.7	815.7
20.5	15.2	345.3
20.6	40.3	875.8
17.6	10.1	890.2
56.3	20.8	710.4
1112.2	392.2	907.4
806.3	1035.8	1285.4
514.5	434.8	908.5
25.1	5.5	490.1
10.4	40.2	450.8
40.7	5.8	1175.1
5.2	15.1	1085.3
30.3	25.6	485.5
25.6	40.5	395.8
40.8	25.8	665.5

Source: D. Gupta, 2017

When Size Matters ….

Padding with 0's

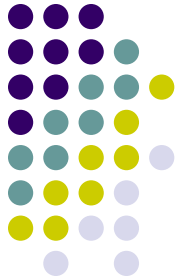
How does that address the issues?



0	86	4	8	184	0	6.4	63.2	184		
0	252	3	8	40	0	75.6	252.9	5.4	20	40
0	34	7	7	163	0	10.2	36.1	9.1	55.9	163
0	105	2	3	69	0	31.5	105.6	2.9	23.7	69
0	56	3	8	175	0	16.8	56.9	5.4	60.5	175
0	126	1	2	178	0	37.8	126.3	1.6	55.4	178
0	163	8	4	142	0	48.9	165.4	9.2	46.6	142
0	22	222	74	180	0	6.6	88.6	244.2	128	180
0	163	158	204	253	0	48.9	210.4	219.2	279.9	253
0	245	98	85	180	0	73.5	274.4	123.5	139	180
0	1	5	1	98	0	0.3	2.5	5.3	30.4	98
0	4	2	8	90	0	1.2	4.6	4.4	35	90
0	7	8	1	235	0	2.1	9.4	8.3	71.5	235
0	2	1	3	217	0	0.6	2.3	1.9	68.1	217
0	3	6	5	97	0	0.9	4.8	7.5	34.1	97
0	6	5	8	79	0	1.8	7.5	7.4	31.7	79
0	8	8	5	133	0	2.4	10.4	9.5	44.9	133

Source: D. Gupta, 2017

Preserving the Spatial Arrangement of Pixels



Weight matrix, aka “filter”

86	4	8	184
252	3	8	40
34	7	7	163
105	2	3	69
56	3	8	175
126	1	2	178
163	8	4	142
22	222	74	180
163	158	204	253
245	98	85	180
1	5	1	98
4	2	8	90
7	8	1	235
2	1	3	217
3	6	5	97
6	5	8	79
8	8	5	133

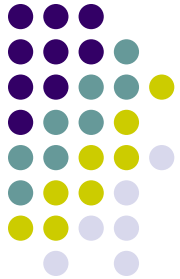
1	0.3
0.5	2

=SUMPRODUCT(B2:C3,$G$5:$H$6)

147.2		
283.9	22.9	349.5
92.6	16.1	195.4
139.6	20.4	377.7
121.9	9.9	417.5
223.8	13.6	341.4
620.4	268.2	443.6
486.1	731.2	736
528.9	438.2	682.4
284.9	128	335.5
8.5	22.3	214.4
24.1	10.4	505.5
12.4	14.8	507
15.8	14.9	264.6
17.8	26	196.1
27.5	21.4	300.2

Source: D. Gupta, 2017

Weight (or Filter) “Stride”



Values of weight matrix chosen through learning

But what is stride?

Why is it important?

INPUT IMAGE

18	54	51	239	244	188
55	121	75	78	95	88
35	24	204	113	109	221
3	154	104	235	25	130
15	253	225	159	78	233
68	85	180	214	245	0

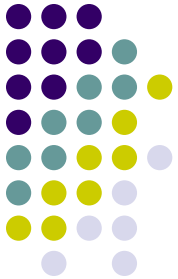
WEIGHT

1	0	1
0	1	0
1	0	1

429	505	686	856
261	792	412	640
633	653	851	751
608	913	713	657

Source: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Stride & Processed Image Size



INPUT IMAGE

18	54	51	239	244
55	121	75	78	95
35	24	204	113	109
3	154	104	235	25
15	253	225	159	78

WEIGHT

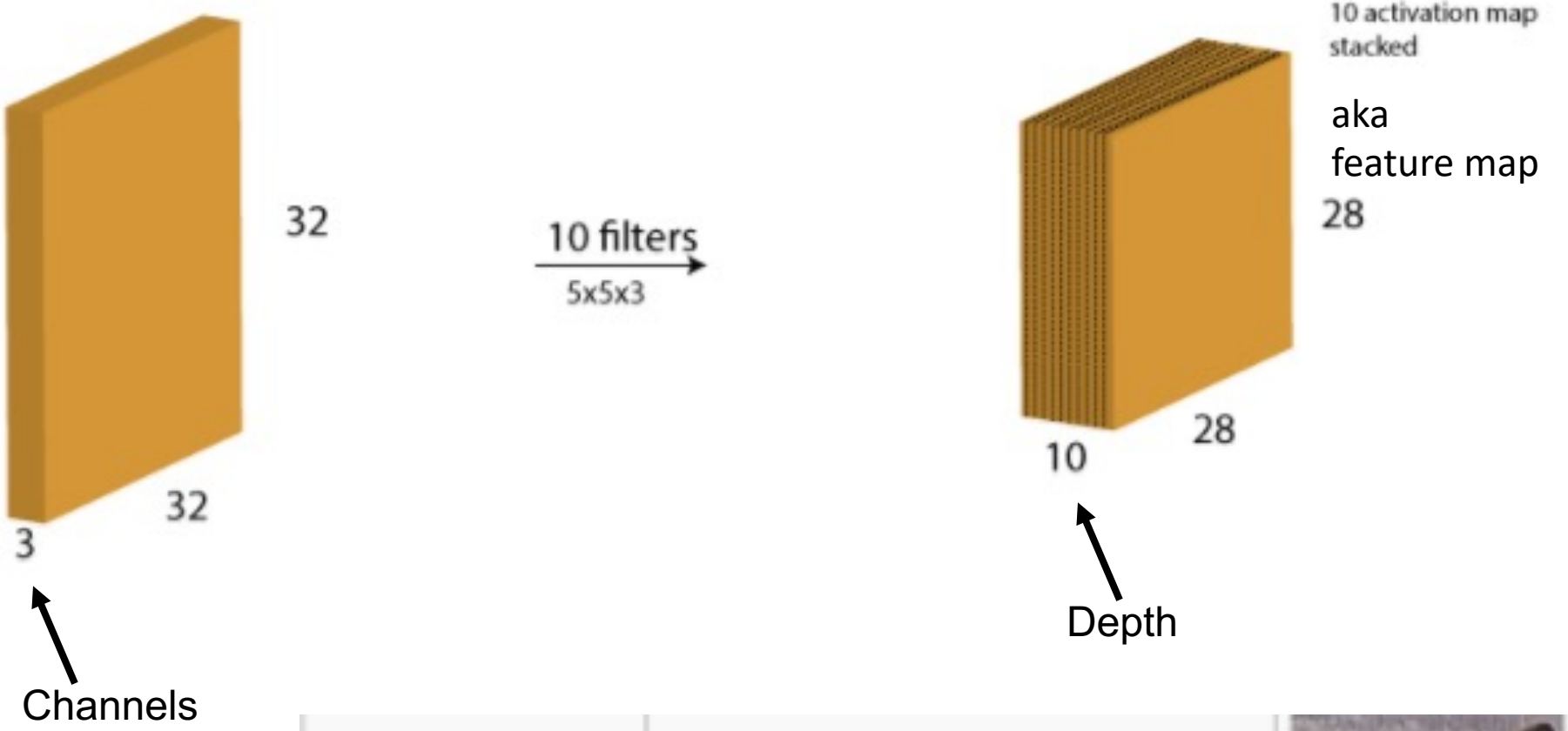
1	0	1
0	1	0
1	0	1

429	686
633	412

What is the stride in the above convolution?
What can be the issue with larger strides?
What would be a solution?

Source: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Summary of the Conv. Layer

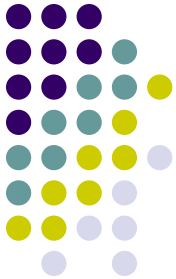


Multiple filters may perform different functions

Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	

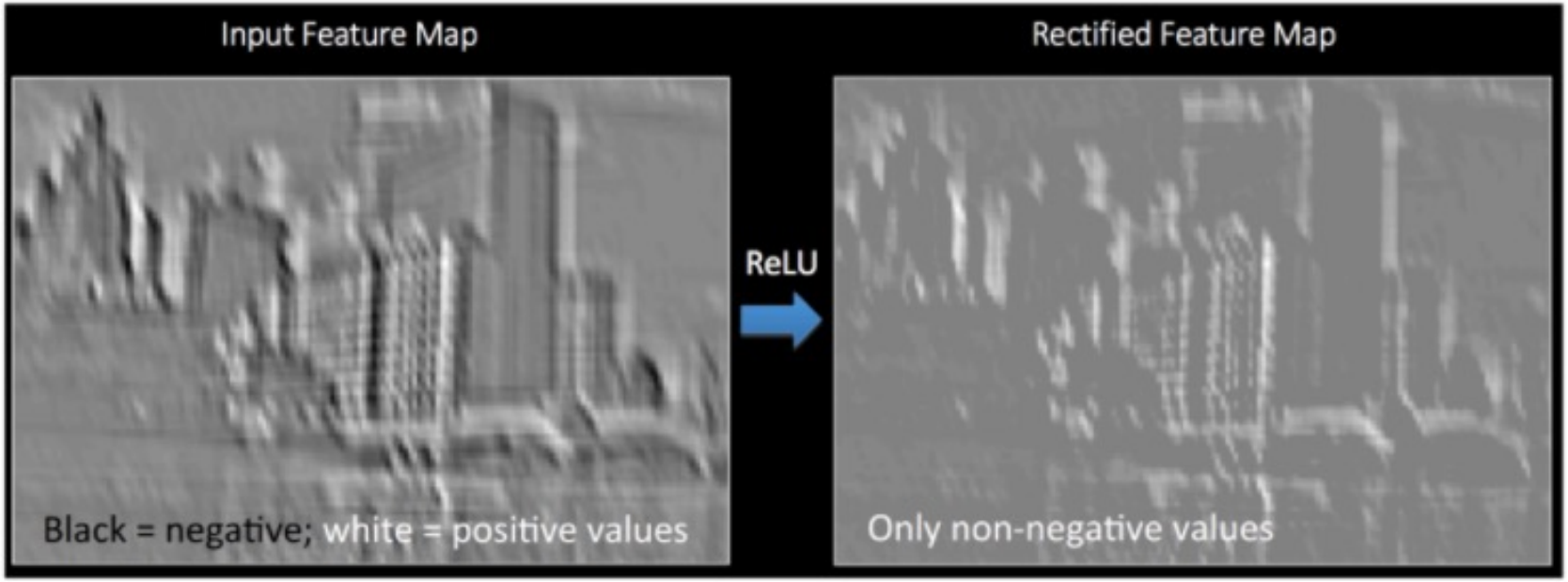
Source: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Bringing Non-linearity (ReLU)



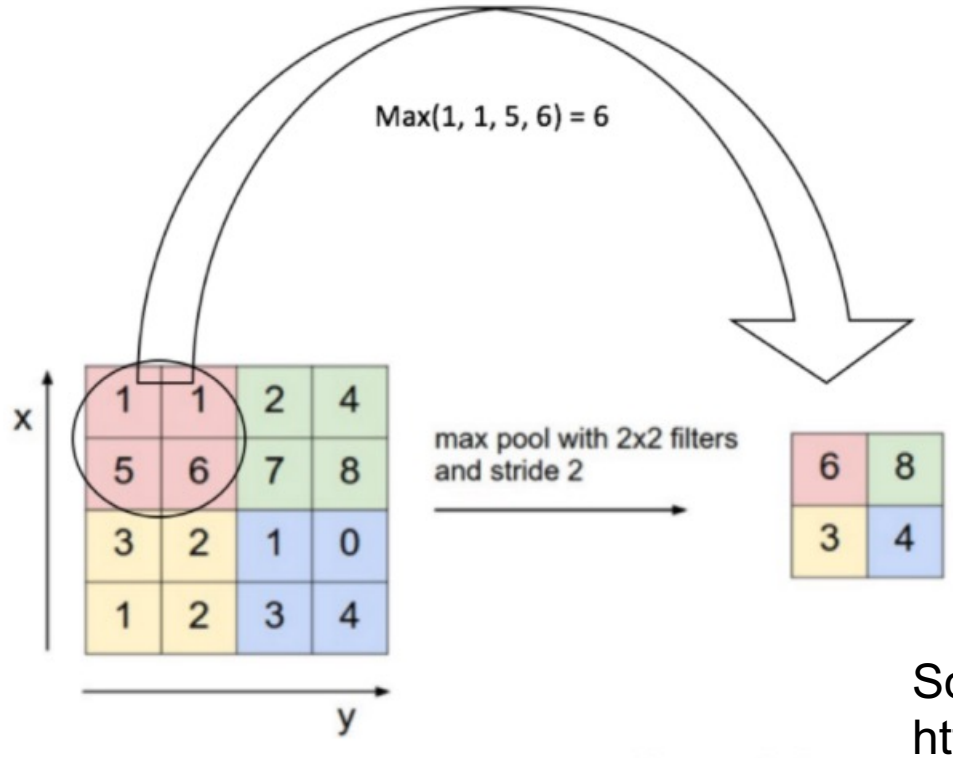
Rectified Linear Unit (ReLU)

Why bother for images?

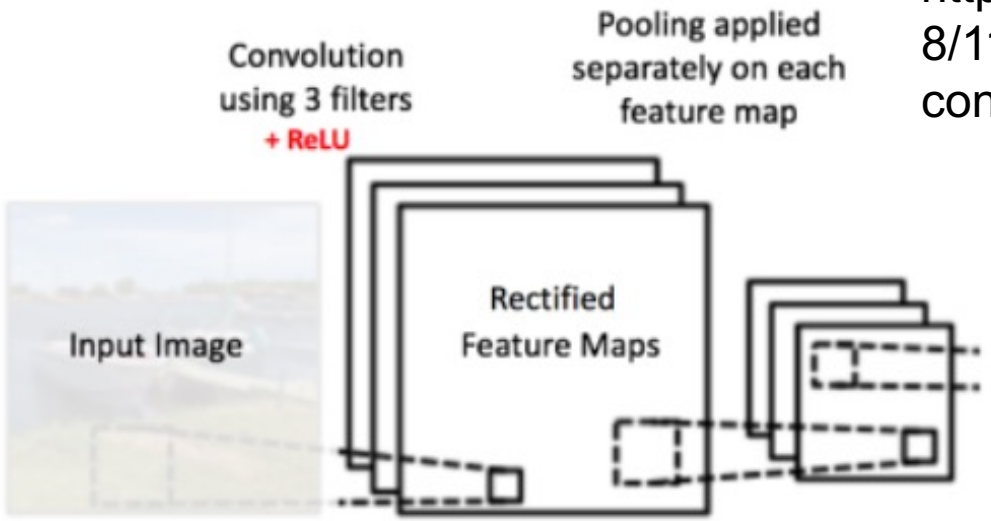


Source: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

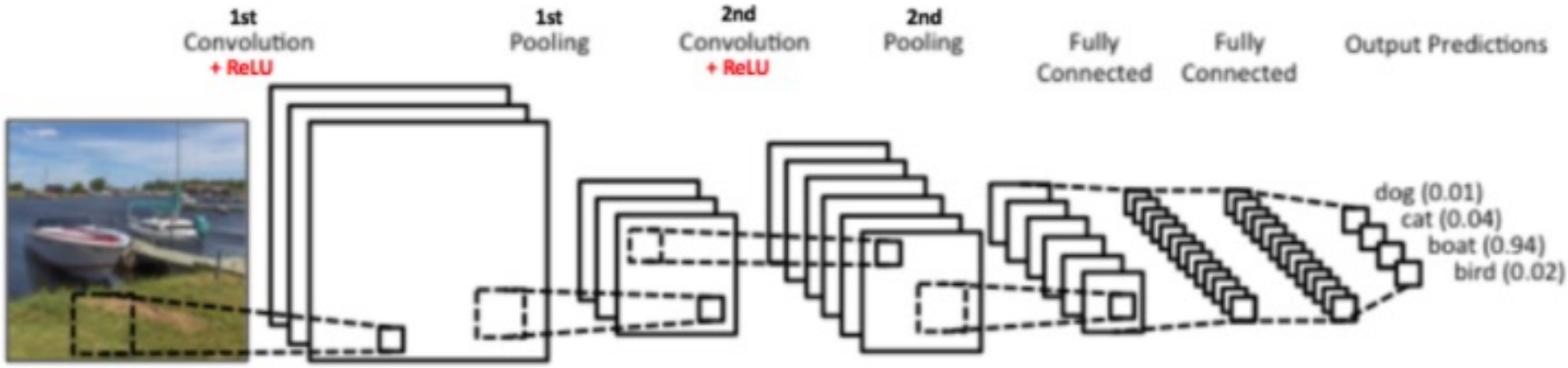
Pooling (Downsampling)



Source: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/



The Big Picture of a ConvNet (CNN)



Check out http://scs.ryerson.ca/~aharley/vis/conv/flat.html for a demo

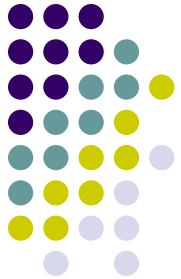
Source: https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Network Analytics Part I

MSITM, October 17, 2022

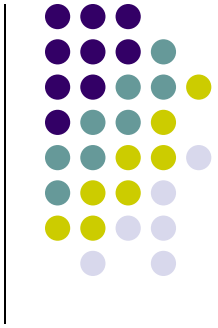
Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor of Business
Distinguished Fellow, INFORMS Information Systems Society
University of Texas Distinguished Teaching Professor
Associate Director, Center for Research in e-Commerce
McCombs School of Business, University of Texas at Austin
Email: aniteshb@gmail.com

Focusing on Connections



- From content to connections
  - Extracting insights from a networked world
  - Why connections matter
  - What we can predict from connections
- Concepts are common to any kind of network: Social, professional, internal corporate, e-commerce, etc.
- Not just humans: Relationships between products, diseases, stocks, etc.

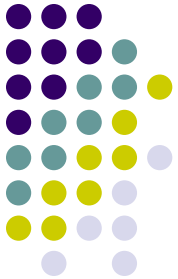
A Networked View of Stocks



	s_1	s_2	….	s_n
s_1	1	c_{12}	…	c_{1n}
s_2		1	…	c_{2n}
…			1…1	…
s_n				1

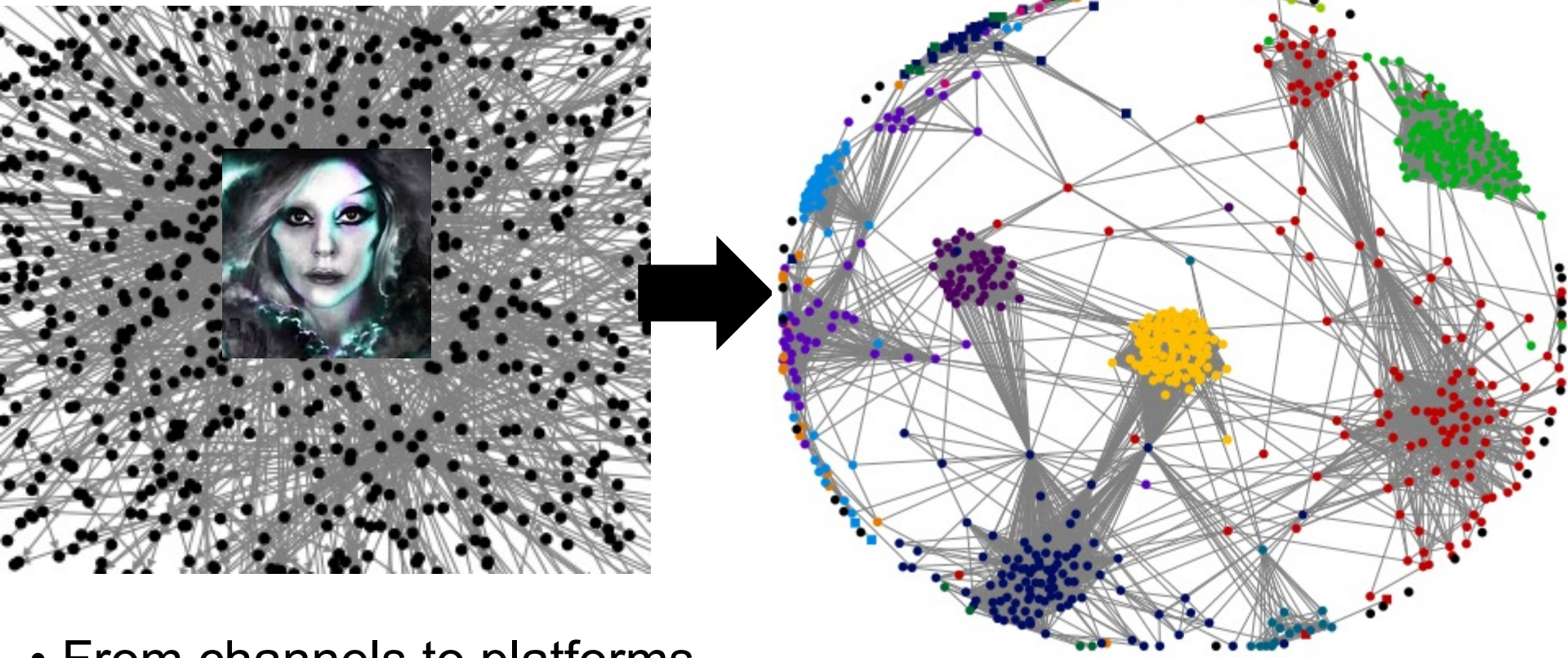
- Stocks s_i and s_j ($i \neq j$) will have a link (edge) if price correlation $c_{ij} \geq$ threshold.
- Many important properties can be studied using network analytics
- E.g., which stock is most important in explaining price movements of the group?
- Other ways to create a network of stocks?

Topics to be Covered



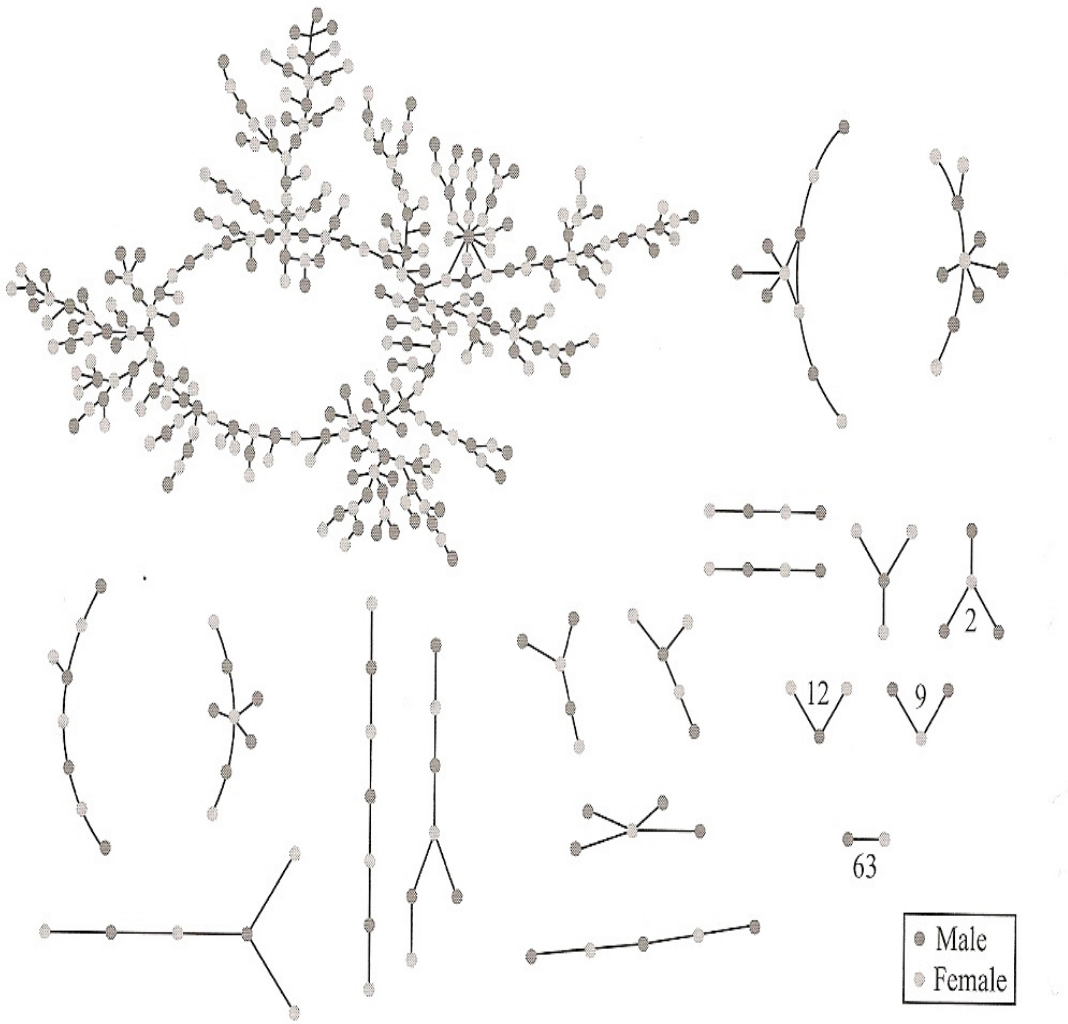
- Unique aspects of social media
- Who's important: Attention & influence
- Visualization with networks
- Detecting communities
- Multi-mode networks
- Homophily vs. social influence
- Network value of a customer

What's Unique About Social Networks?
1. "Democratization of the Lady Gaga Effect"



- From channels to platforms
  - Conversations are now visible
  - Connections, attention, influence, etc. can be measured

2. Network Structure

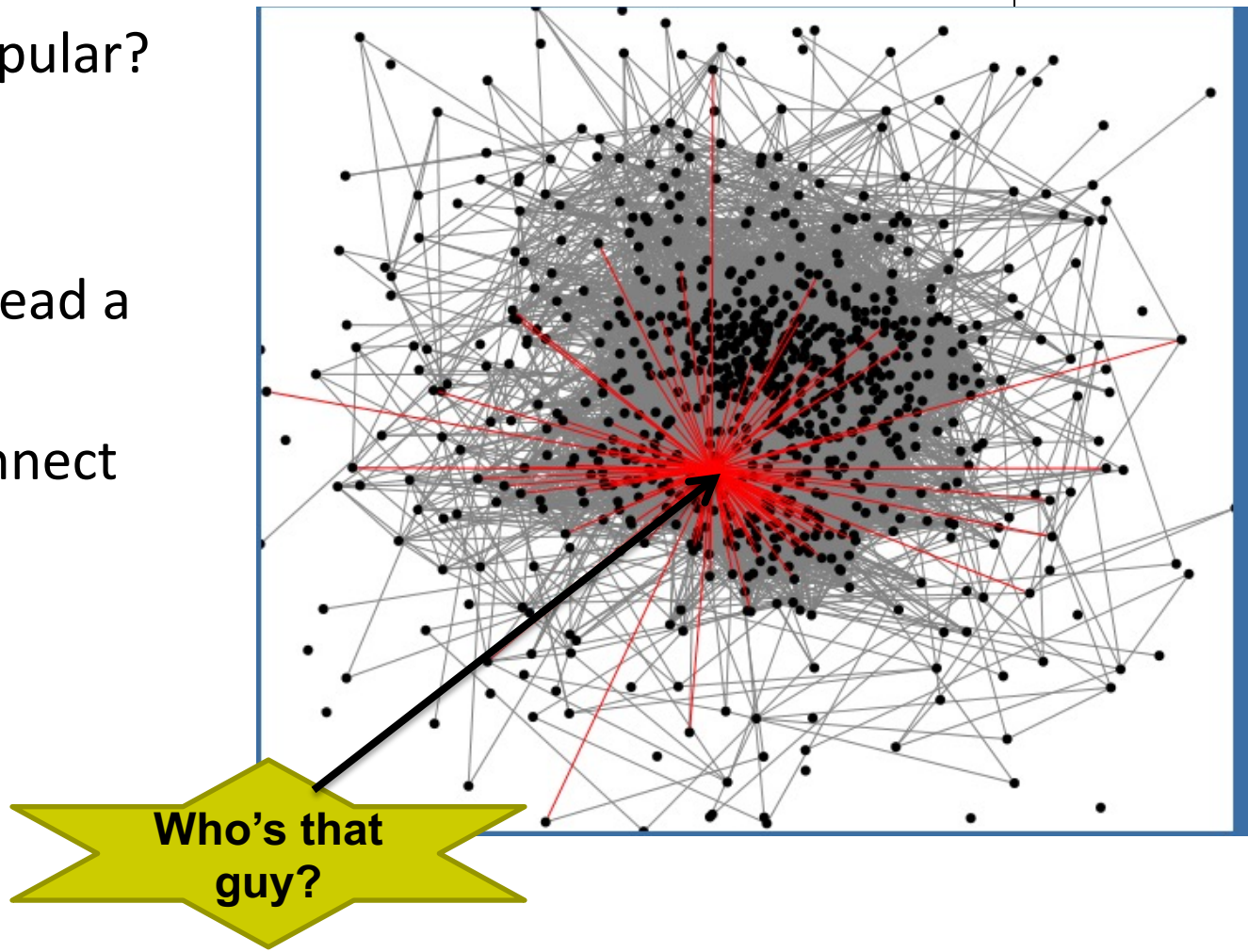


Can we learn from epidemiology?

Source: “Networks, Crowds and Markets”

Your Network Location Matters

- Who are most popular?
- Who are “best” connected?
- Who can help spread a message?
- Who can help connect diverse groups?

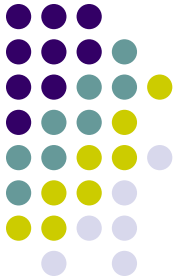


3. The Network Value of a Customer

- From Customer Lifetime Value (*CLV*) to Customer Influence Value (*CIV*)
- Customer Network Lifetime Value (*CNLV*) = *CLV* + *CIV*

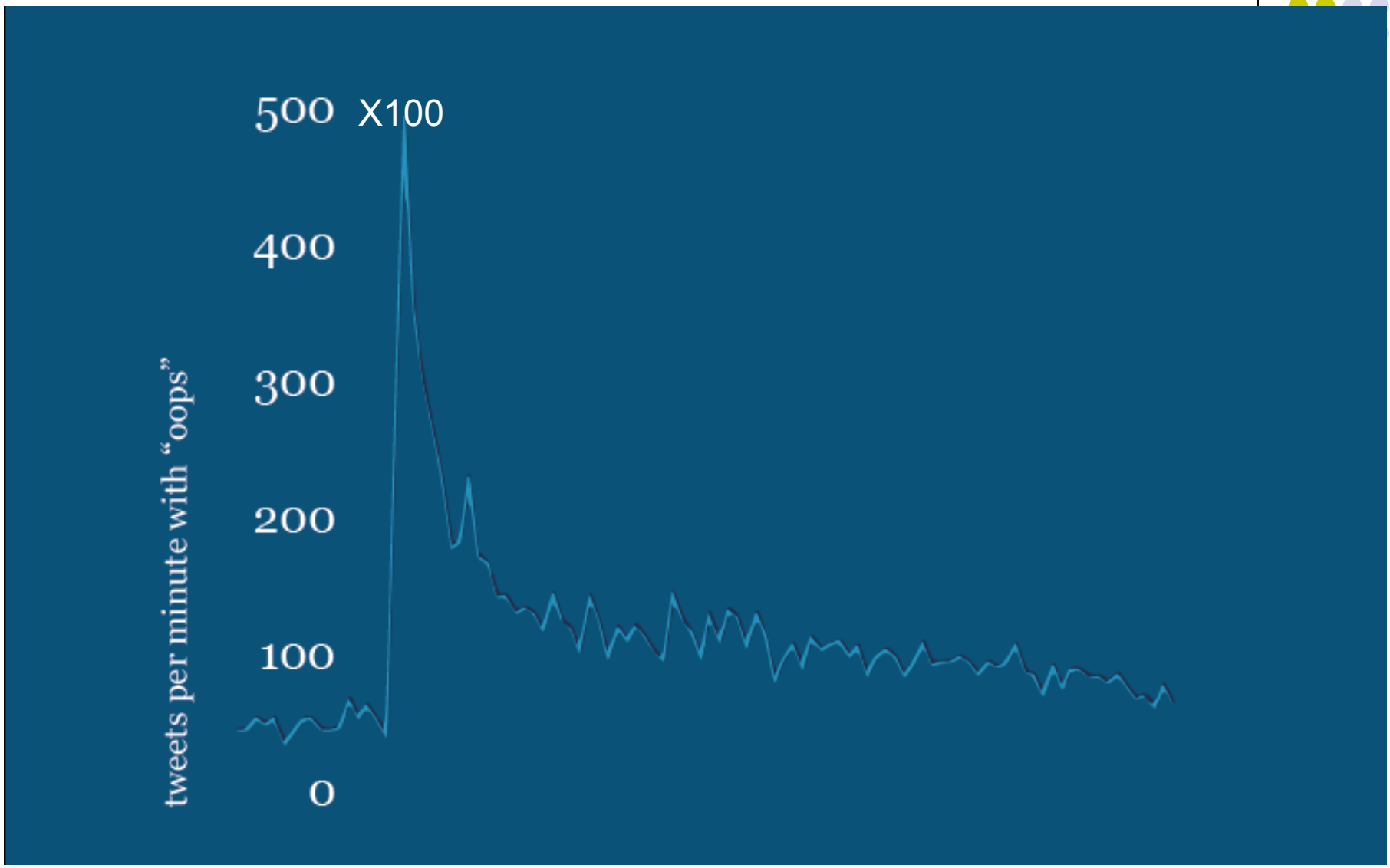
Image Source: http://www.wired.co.uk/news/archive/2012-08/13/customer-network-lifetime-value

4. Monitoring & Visualization Through Social Media



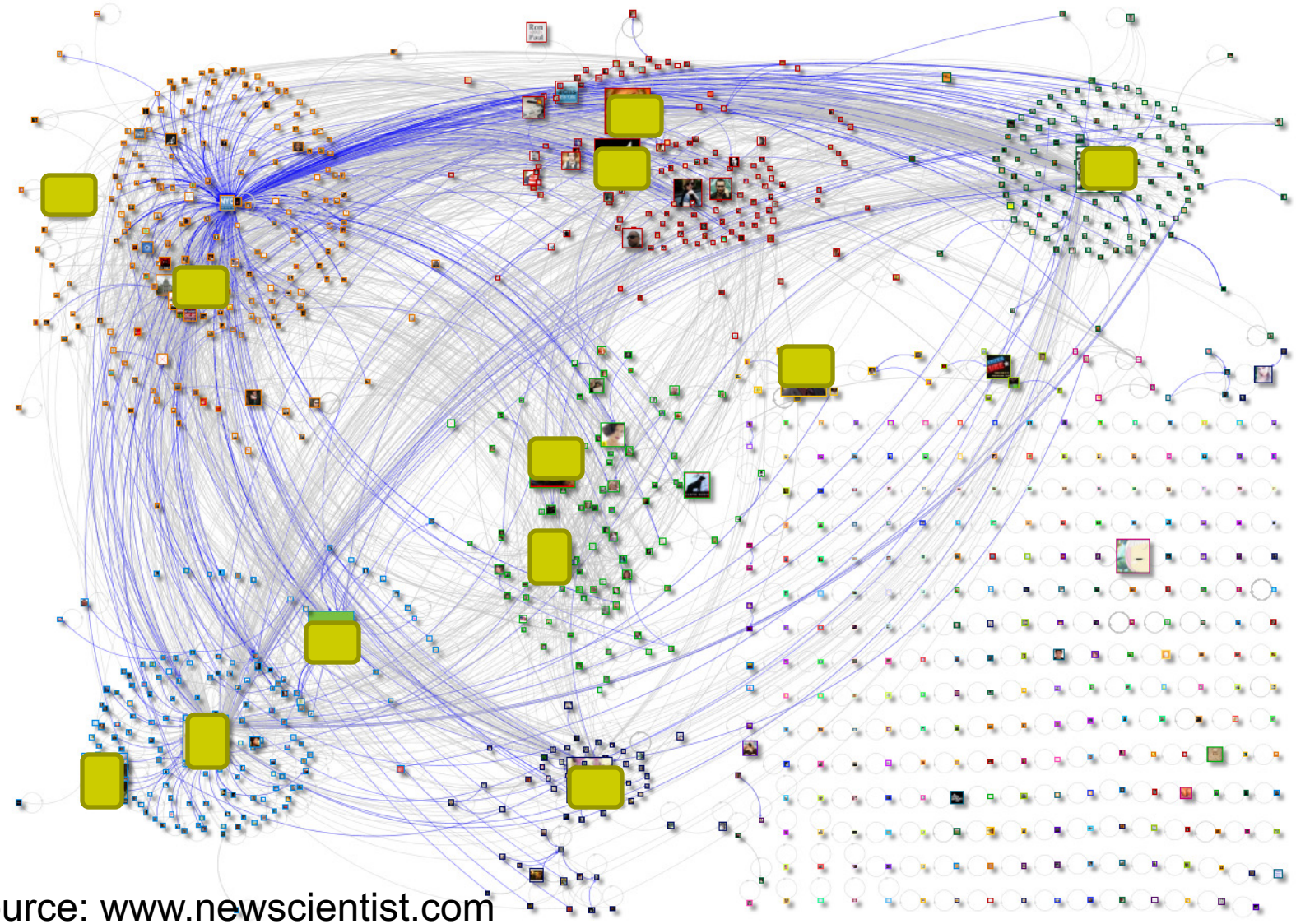
- Unprecedented real time visibility into
  - Public reactions
  - Emerging phenomena or events
  - Customer preferences

Real-time Assessment of Sentiment & Opinion: The “Oops” Tweets



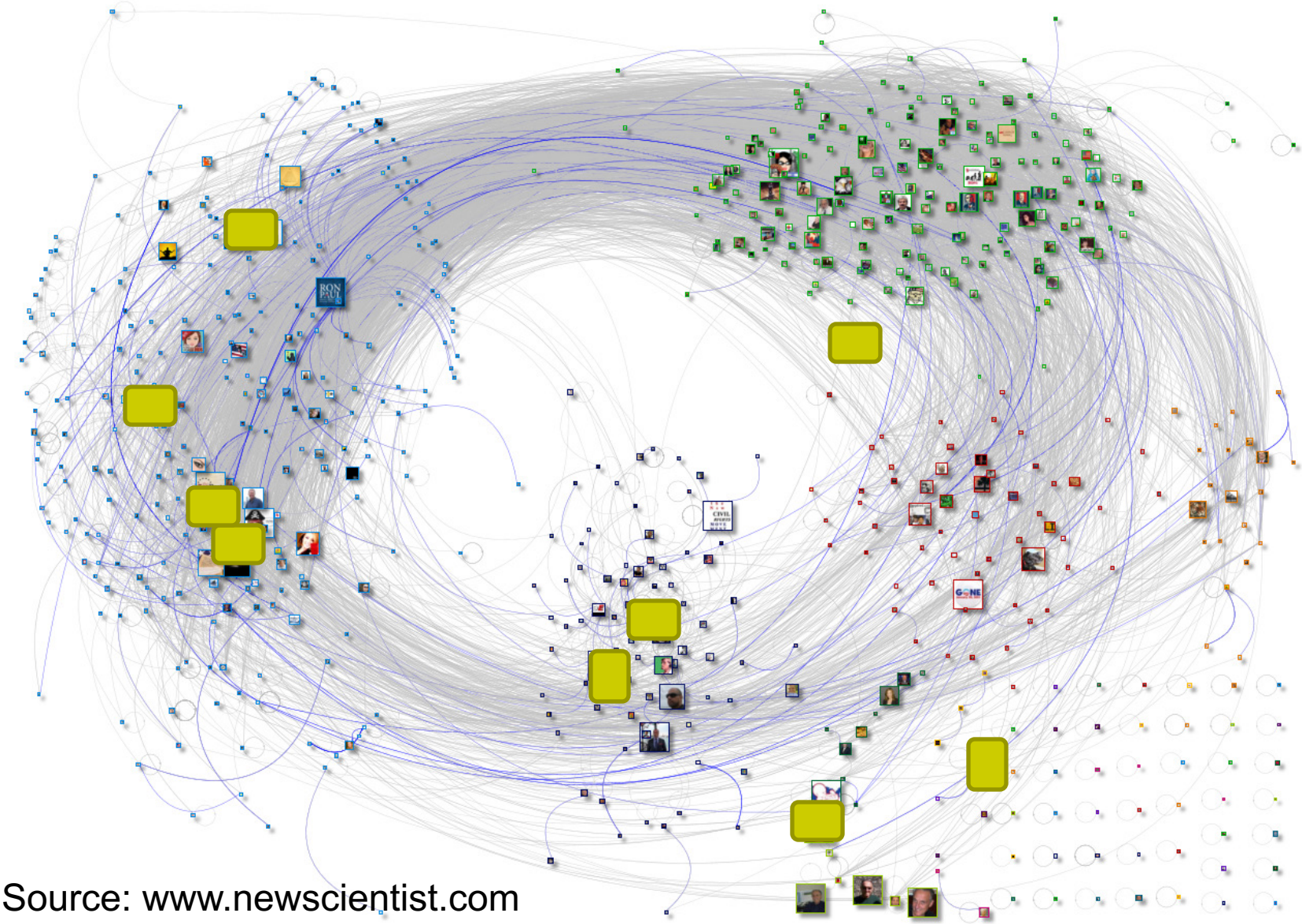
Source: April Underwood, Twitter

Tweets, Re-tweets: What do They Say?



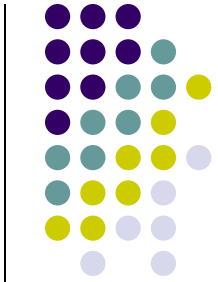
Source: www.newscientist.com

How is This Different From the Previous Network?



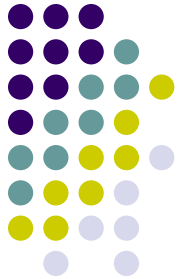
Source: www.newscientist.com

Two Perspectives



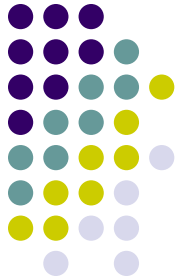
- A social network platform's perspective
  - How to increase # connections & interactions
  - Metrics to track such growth
  - Targeting
- A user organization's perspective
  - Who are important for our brand or product?
  - Metrics to rank them
  - How to pursue them
  - How networked customers make decisions

Describing Your Network to an Advertiser



- What metrics can we use to describe the connectedness of the network?

A Reality Check Before Diving Into Theories

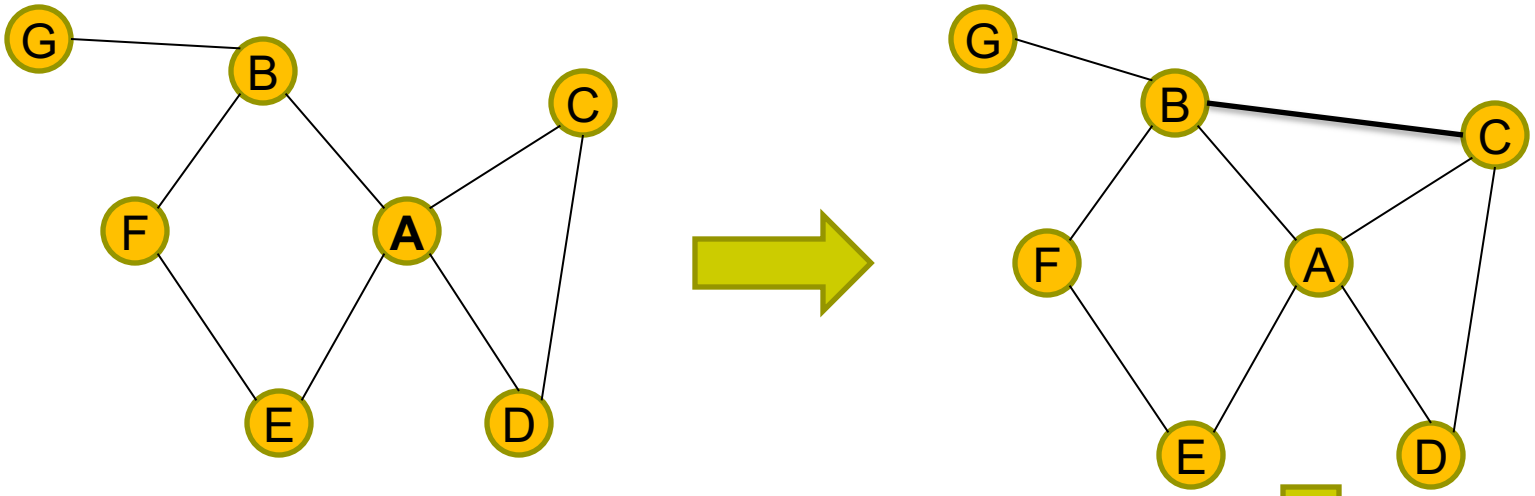


- “If you are friends with Alan, and friends with Betty, then it is likely that Alan and Betty will become friends as well, mostly because they already have something in common: You.”

- “You brought a friend to your favorite yoga studio and she started regularly attending class, even when you didn’t go.”

Source: http://plainspokenlinguist.wordpress.com/2013/09/20/i-know-a-guy-the-power-of-triadic-closure/

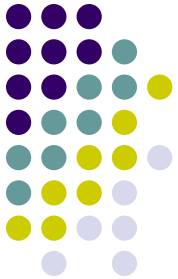
Predicting Future Links (Edges) With "Triadic Closure"



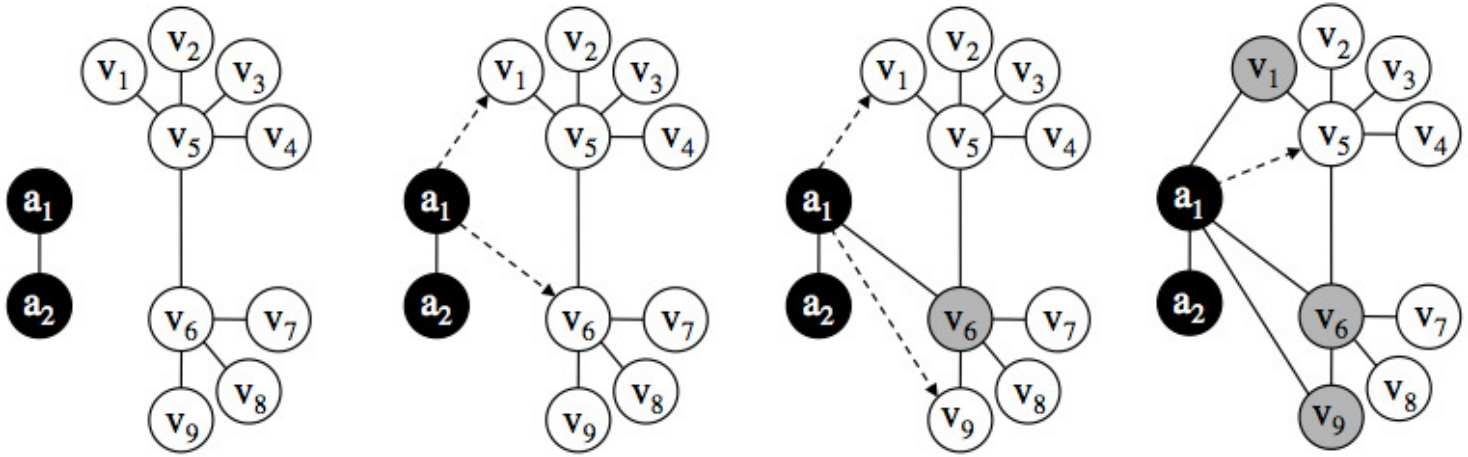
- "If two people in a social network have a common friend, then there is an increased likelihood that they will become friends themselves at some point in the future"
- "People you may know" in FB
- ***Clustering coefficient*** of a user: Probability that two randomly selected friends of the user are friends with each other.

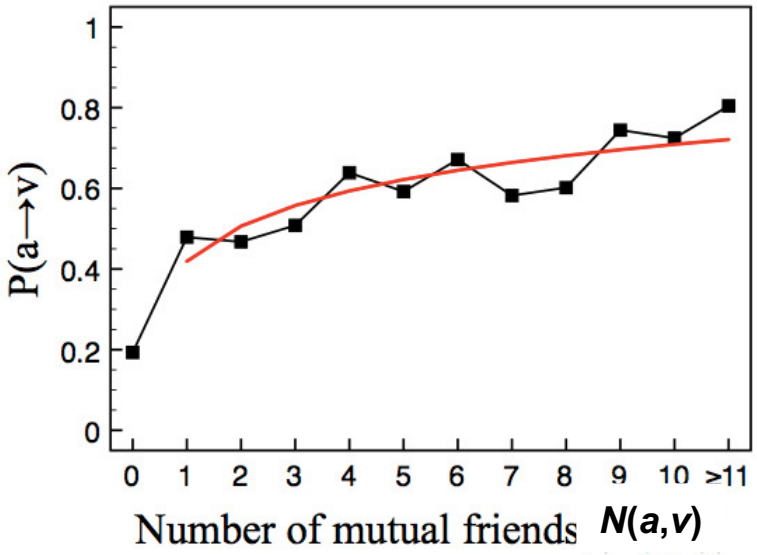
Source: Easley & Kleinberg, "Networks Crowds & Markets"

Abusing the Triadic Closure Principle



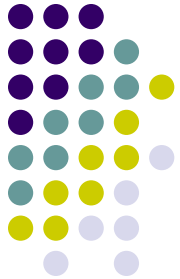
UBC students wrote code that randomly sends friend requests
If accepted, then …?
8,954 users requested, 3,055 accepted





Source: http://blogs.cornell.edu/info2040/2011/09/30/abusing-triadic-closure/

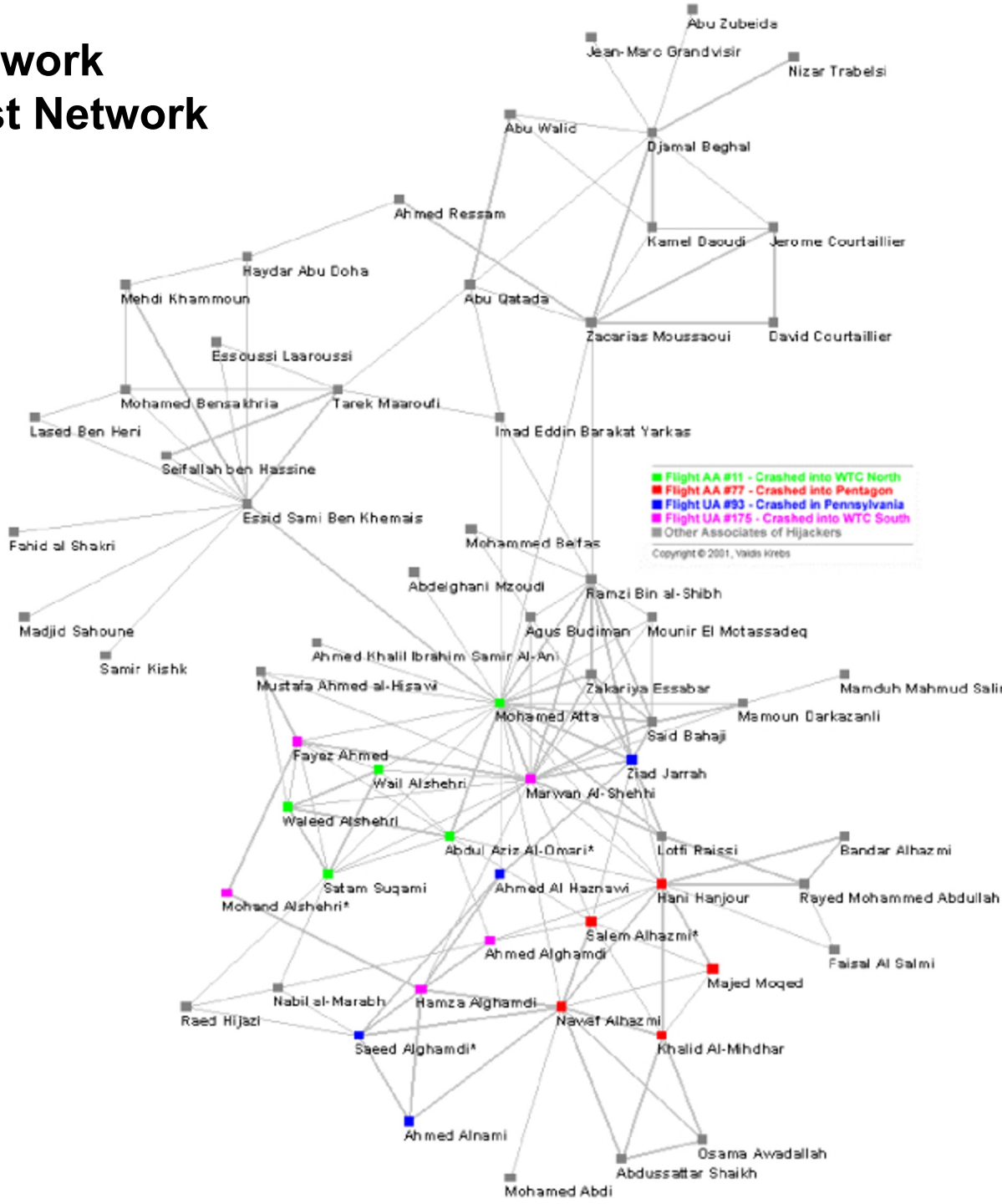
Social Network Structure



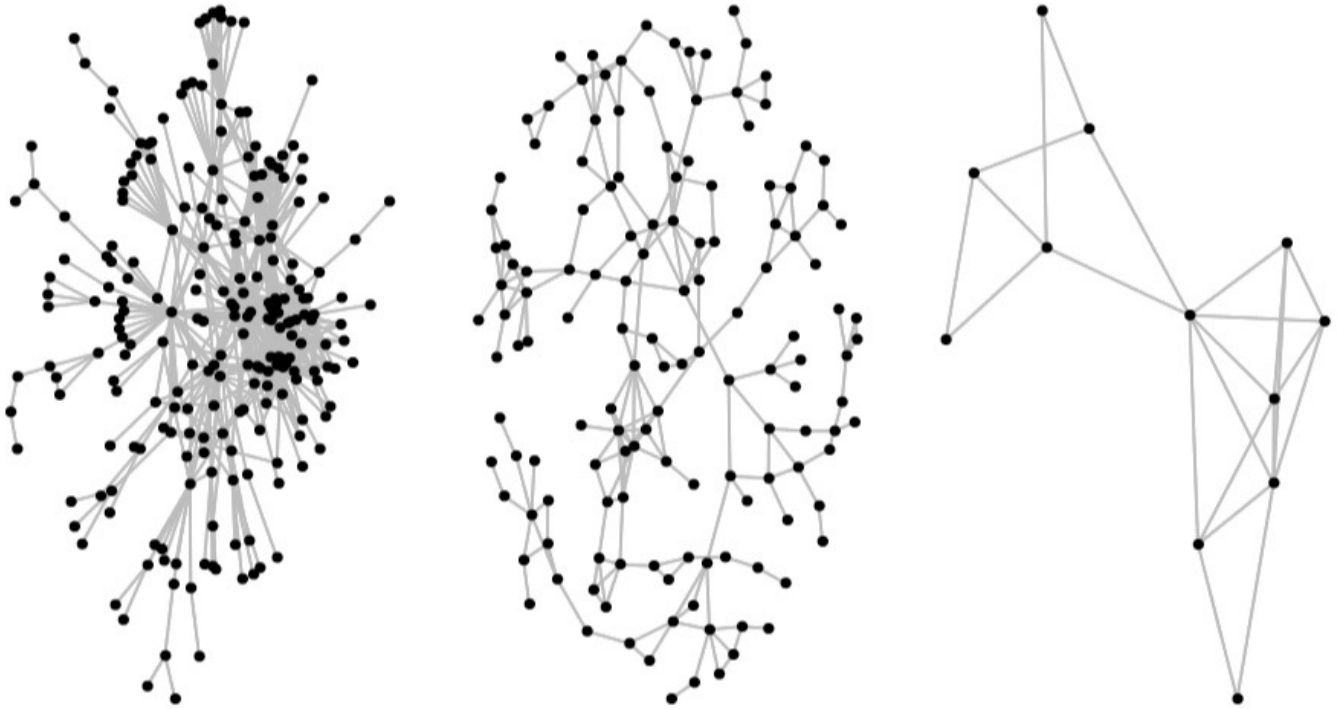
- In addition to getting attention & being active, your position in a network matters
- E.g.,
  - Who are most popular?
  - Who can spread information quickly?
  - Who help connect diverse groups?
- Need to look into the structure of networks

Myriad Applications of Network Analytics: The 911 Terrorist Network

Who are central to the network?
What was the role of M. Atta?
How can we watch out against future attacks?



Not All Networks Are Created Equal

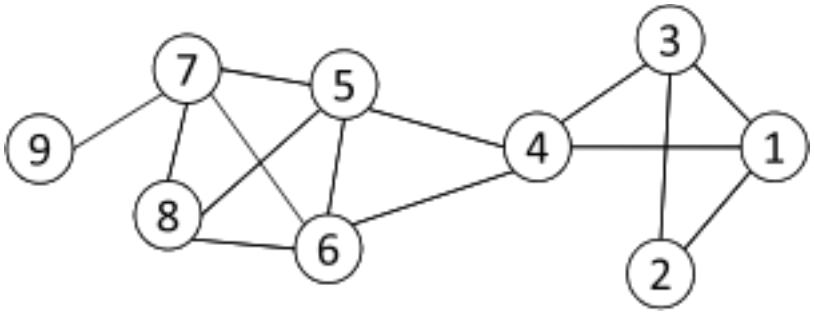


How do we summarize the essential properties of these networks?

Source: P. Hoff, U. Washington

Metrics: Degree Centrality

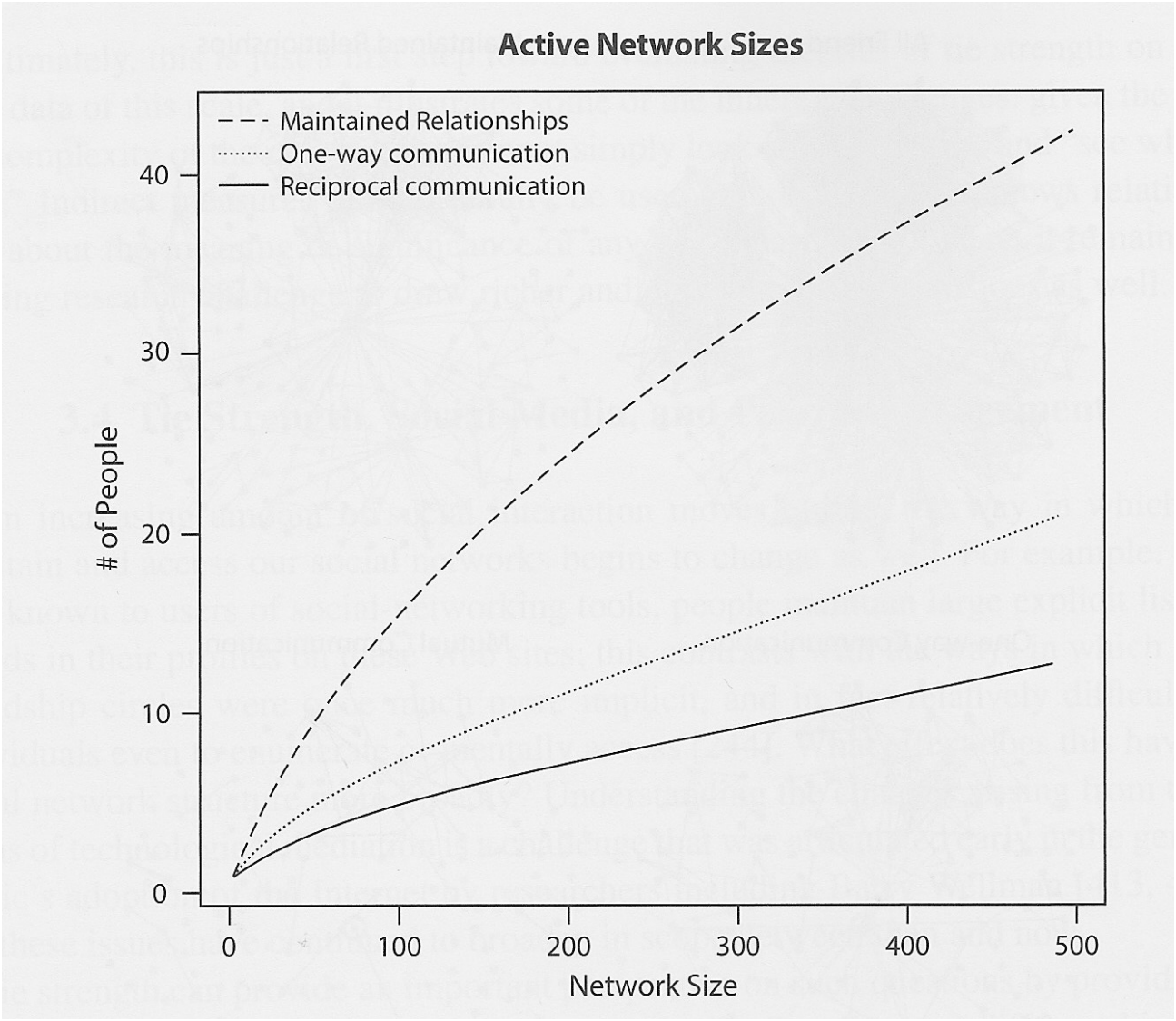
- Who are most popular? Most connected?
- Number of “edges” connected to a “node” or “vertex”
- Normalized Degree Centrality: Degree centrality/(n -1)



For node 1, degree centrality is 3; Normalized degree centrality is 3/(9-1)=3/8.

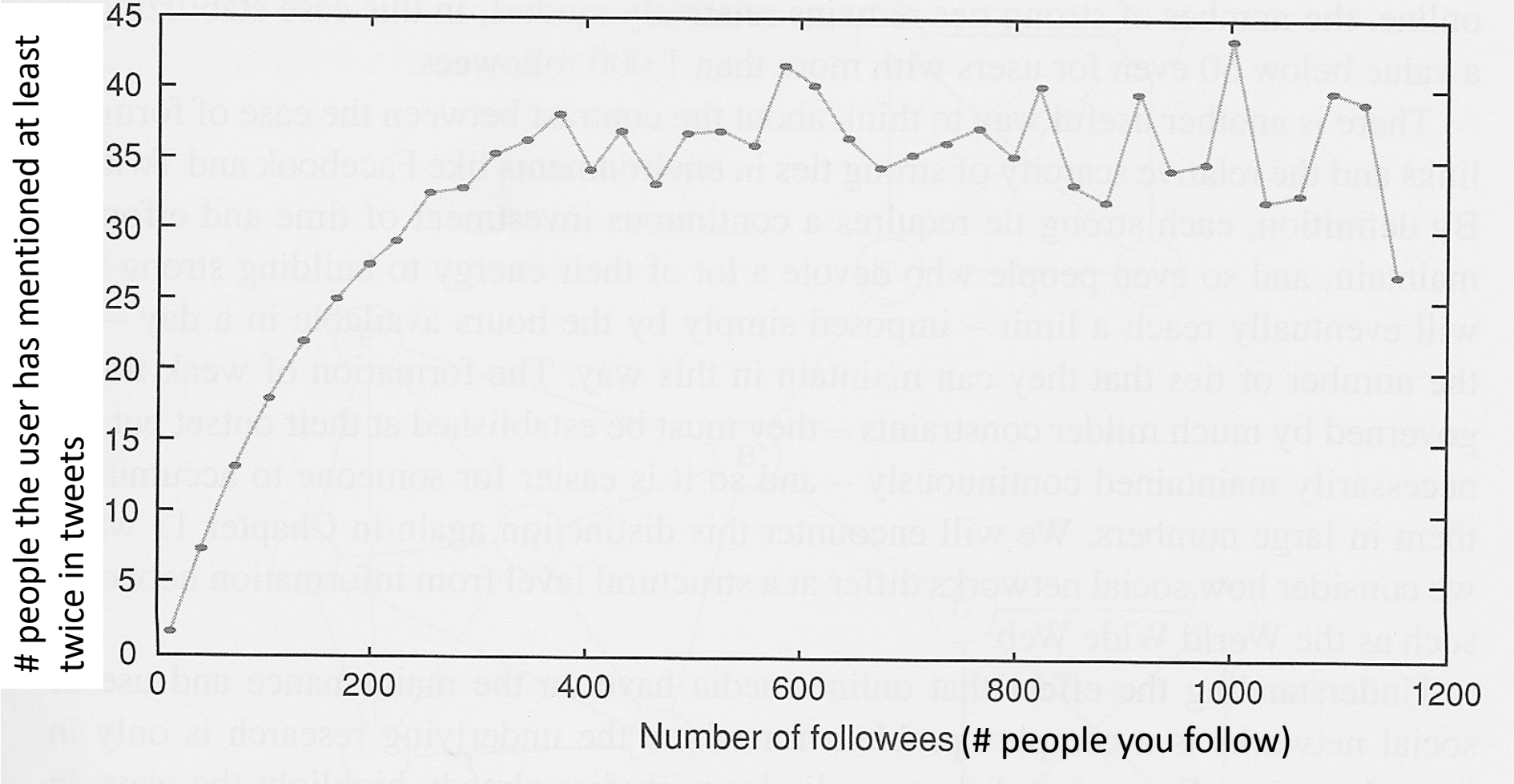
- In-degree and out-degree for directed networks (e.g., Twitter, email, etc.)
- Can degree be a useful metric?

Is Degree a Good Indicator of Activity?



Source: Easley & Kleinberg, “Networks Crowds & Markets”

Strength of Ties on Twitter

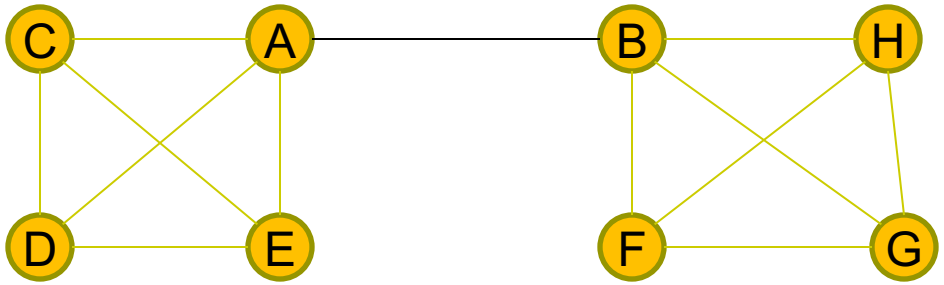


- What can we conclude here?

Source: Easley & Kleinberg, “Networks Crowds & Markets”

"The Strength of Weak Ties"

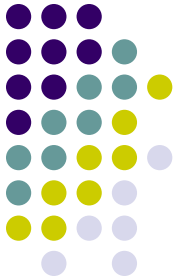
- Granovetter's observations on job leads



- A has four friends, but the friendships are different
- A, C, D and E probably share "strong" ties
- B may belongs to a different, distant world
- A-B possibly represents a "weak" tie
- But may be a source of new information, ideas or insights
- Captured by the "betweenness" centrality metric

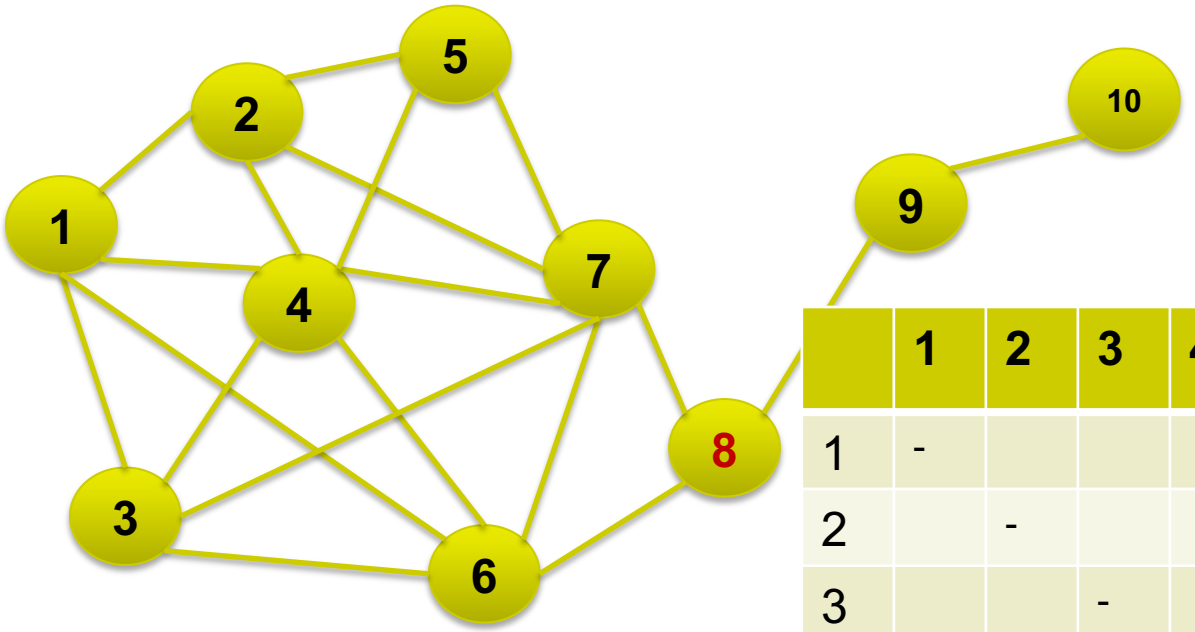
Source: Easley & Kleinberg, "Networks Crowds & Markets"

Network Metric: Betweenness Centrality



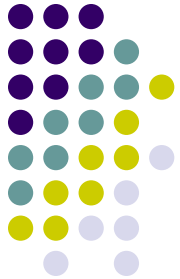
- Popularity (degree) is useful, but there are other roles
- Connecting *disparate* parts of a network
- How difficult will it be for others (esp. at the “outskirts”) to communicate without you in the network?
- Nodes with high betweenness are important in transmitting new information, ideas & opportunities to a wide audience

Betweenness Centrality Example



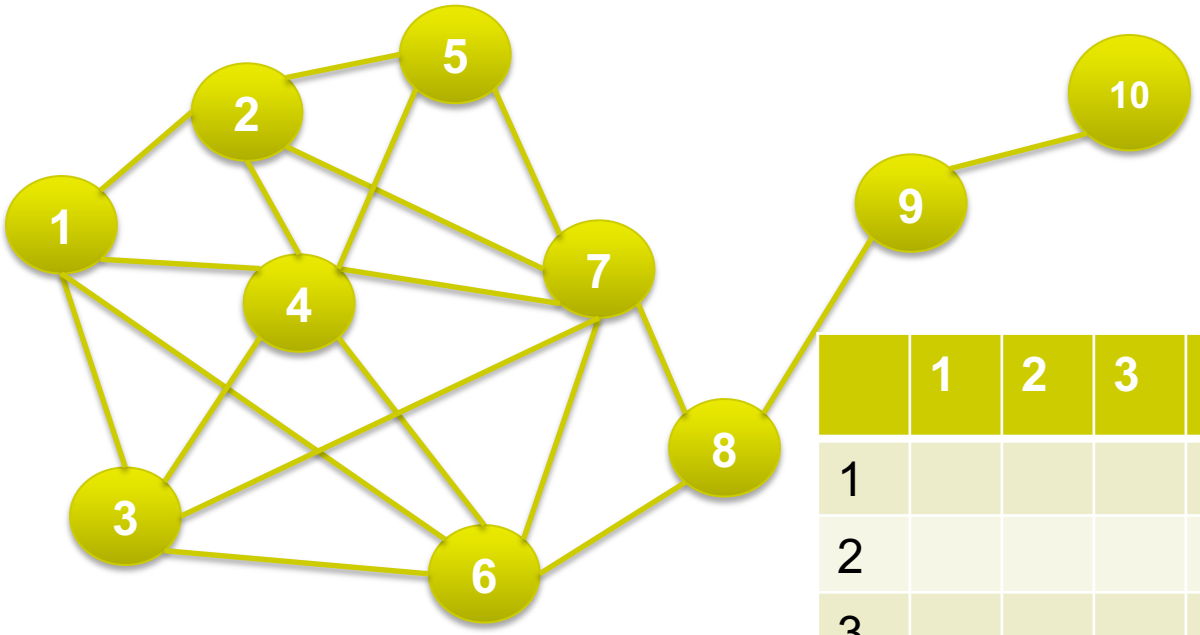
	1	2	3	4	5	6	7	9	10
1	-								
2		-							
3			-						
4				-					
5					-				
6						-			
7							-		
9								-	
10									-

Metric: Closeness Centrality



- Some nodes can reach the whole network more quickly than other nodes
- How close a node is to all other nodes
- Create a matrix of shortest distances (geodesic) between nodes
- Average distance between a node and all other nodes
- 1/average distance is the closeness centrality of the node

Closeness Centrality Example



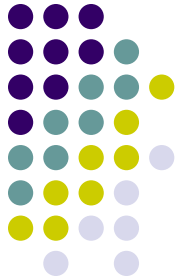
	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

The 911 Terrorist Network Revisited

Who are central to the network?
What was the role of M. Atta?
How can we watch out against future attacks?

Degrees	Betweenness	Closeness
0.417 Mohamed Atta	0.334 Nawaf Alhazmi	0.571 Mohamed Atta
0.389 Marwan AI-Shehhi	0.318 Mohamed Atta	0.537 Nawaf Alhazmi
0.278 Hani Hanjour	0.227 Hani Hanjour	0.507 Hani Hanjour
0.278 Nawaf Alhazmi	0.158 Marwan AI-Shchhi	0.500 Marwan AI-Shehhi
0.278 Ziad Jarrah	0.116 Saeed AIghamdi*	0.480 ZiadJarrah
0.222 Ramzi Bin al-Shibh	0.081 Hamza Alghamdi	0.429 Mustafa al-Hisawi
0.194 Said Bahaji	0.080 WaleedAlshehri	0.429 Salem Alhazmi*
0.167 Hamza Alghamdi	0.076 ZiadJarrah	0.424 Lotfi Raissi
0.167 Saeed Alghamdi*	0.064 Mustafa al-Hisawi	0.424 Saeed AIghamdi*
0.139 Lotfi Raissi	0.049 Abdul AzizAl-Omari*	0.419 Abdul Aziz AI-Omari*
0.128 MEAN	0.046 MEAN	0.393 MEAN

Metric: Eigenvector Centrality



- One's importance is partly determined by "the company one keeps"
- If one has many important friends, s/he should be important ☺
- Eigenvector centrality considers not only your degree, but your friends' degree
  - I.e., are your friends connected to large networks?
- Google's PageRank very similar to Eigenvector centrality

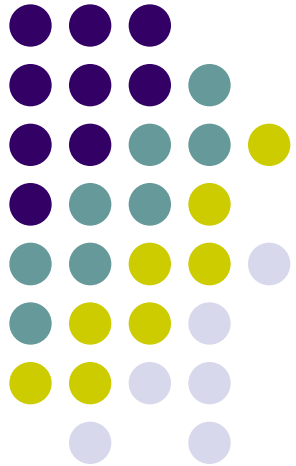
Document Similarity & Applications
Word Embeddings

Unstructured Data Analytics

MSITM, Fall 2022
19th September

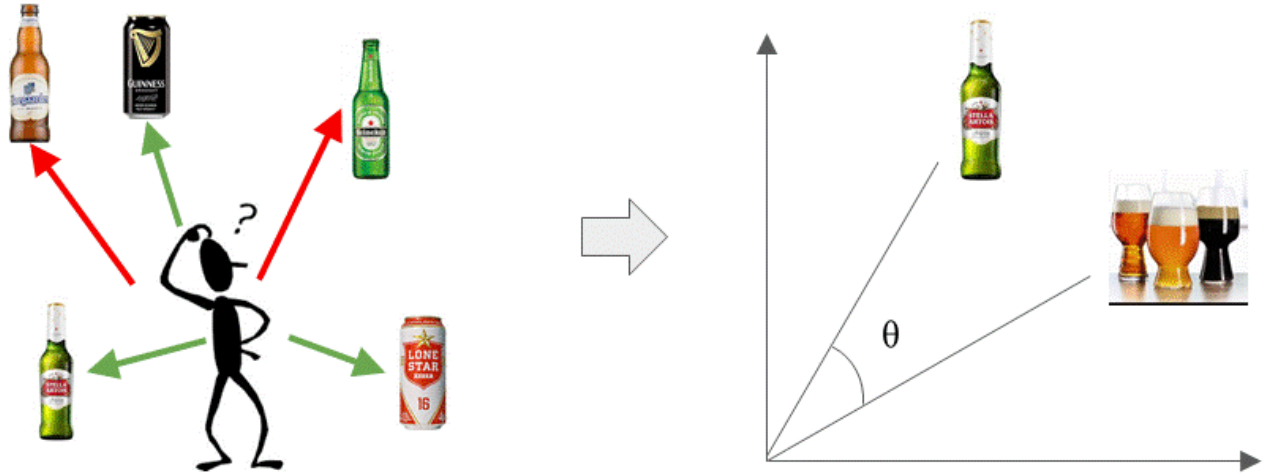
Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor in Business
Distinguished Fellow, INFORMS Information Systems Society
University of Texas Distinguished Teaching Professor
Associate Director, Center for Research in e-Commerce
McCombs School of Business, University of Texas at Austin
Email: **aniteshb@gmail.com**

Document Similarity & Applications
(i) Resonance Analysis
(ii) Crowdsourced Recommendation Systems



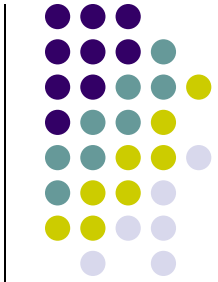
Assessing Document Similarity

- Early applications: Document clustering/categorization & retrieval
- Many business applications
- E.g., which beers or movies are similar?



- Useful for recommender systems – both conventional & crowdsourced
- Understanding your competition in markets with lots of products
- Myriad applications: Did a campaign resonate with the intended audience?

The Basic Idea



- Start with a numerical representation of each document
- Many ways to represent a document with numbers
  - Presence absence of each term (0/1)
  - Raw or scaled (normalized) frequency of each term
  - TF-IDF of each term in a document (be careful, not applicable everywhere)
- Multiple ways to calculate similarity between the documents
- Choice 1: Calculate Euclidean distances between documents
- Choice 2: Calculate the % words that are common across the documents
- Choice 3: Calculate the angle between the two document vectors (many variations are possible)
- Choice 4: Use word embeddings (e.g., word2vec) and calculate similarity

Jaccard Similarity

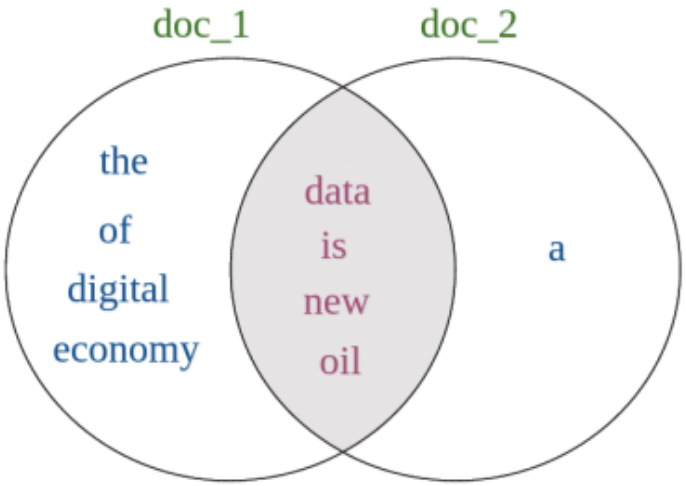
$$Jaccard(doc1, doc2) = \frac{|doc1 \cap doc2|}{|doc1 \cup doc2|} = \frac{|doc1 \cap doc2|}{|doc1| + |doc2| - |doc1 \cap doc2|}$$

Doc1: Data is the new oil of the digital economy
Doc2: Data is a new oil

$$J(doc_1, doc_2) = \frac{\{'data', 'is', 'the', 'new', 'oil', 'of', 'digital', 'economy'\} \cap \{'data', 'is', 'a', 'new', 'oil'\}}{\{'data', 'is', 'the', 'new', 'oil', 'of', 'digital', 'economy'\} \cup \{'data', 'is', 'a', 'new', 'oil'\}}$$

$$= \frac{\{'data', 'is', 'new', 'oil'\}}{\{'data', 'a', 'of', 'is', 'economy', 'the', 'new', 'digital', 'oil'\}}$$

$$= \frac{4}{9} = 0.444$$

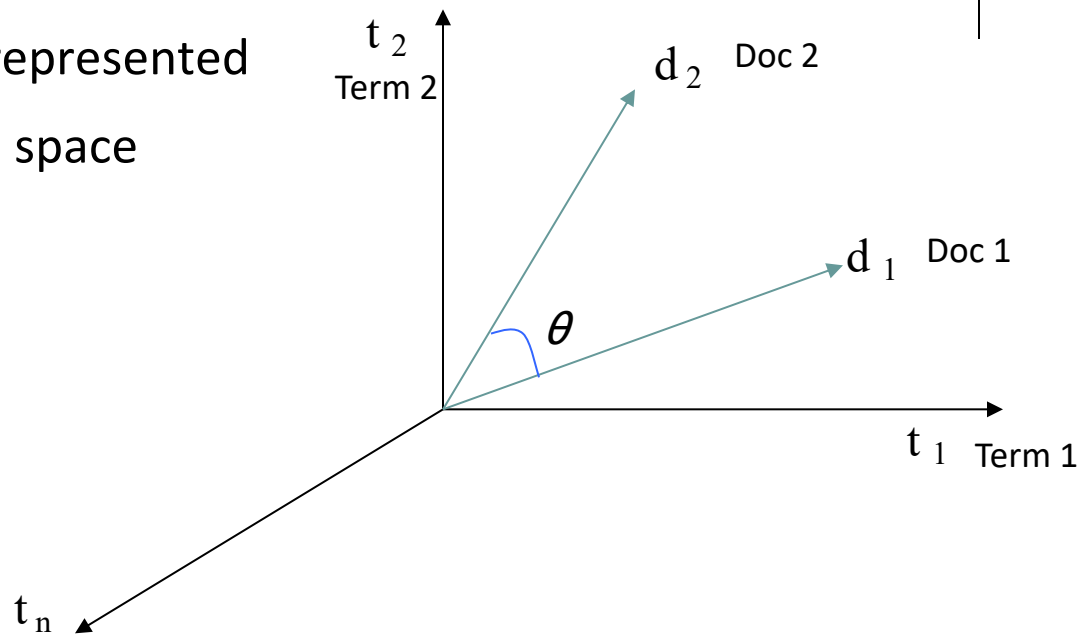


Source: B. Kanani

Angles Instead of Distances

- *n*-dimensional space with *n* terms
- A document (e.g., tweet) represented

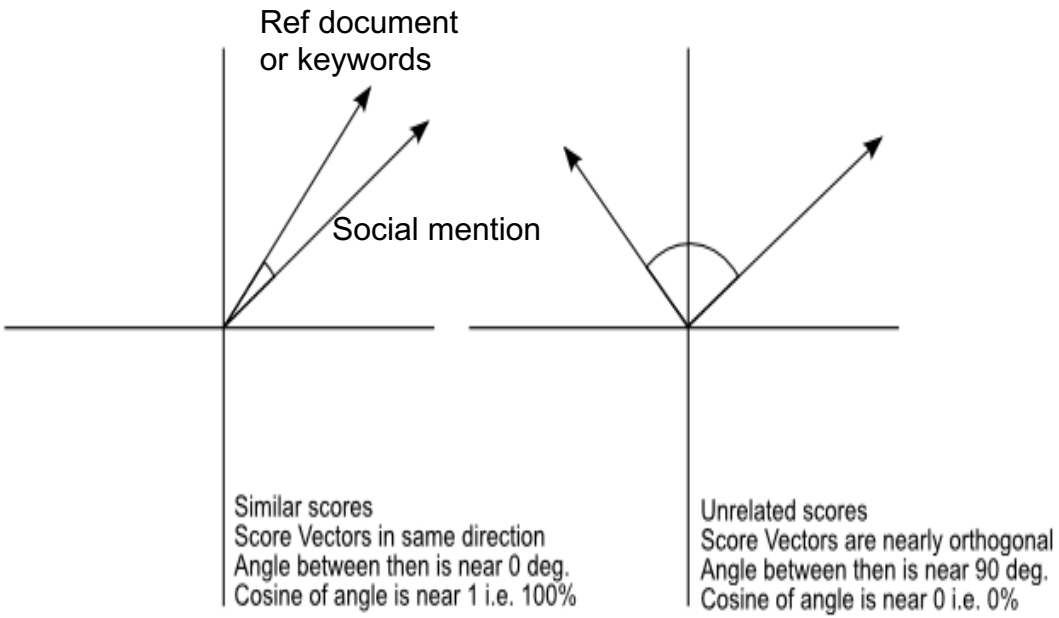
by its coordinates in this *term* space



- Similarity between documents d_1 and d_2 (could also be a set of keywords) is the cosine of the angle Θ between them.
- Score between 0 and 1.

Source: http://pyevolve.sourceforge.net/wordpress/?p=2497

Similar vs. Unrelated Mentions



Is this approach better than Euclidean distances?

Source: http://pyevolve.sourceforge.net/wordpress/?p=2497

Basics Revisited…

- Dot product of two vectors

$$\vec{a}.\vec{b} = \sum_{i=1}^{n} a_i\, b_i$$

- *n* features, a_i and b_i can be *tf*, *tf-idf*, or other scores depending on application
- What does a dot product mean geometrically?

$\vec{a}.\vec{b} = ||\vec{a}|| * ||\vec{b}|| * cos\theta$ where $||\vec{a}|| = \sqrt[2]{\sum_{i=1}^{n} a_i^2}$ and $||\vec{b}|| = \sqrt[2]{\sum_{i=1}^{n} b_i^2}$

$$\text{Or } cos\theta = \frac{\vec{a}.\vec{b}}{||\vec{a}|| * ||\vec{b}||}$$

In Practical Terms ….

$$\cos\theta = sim(d_j, d_k) = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$

- Cosine of angle between two vectors represented by documents d_j and d_k.
- The denominator involves the lengths of the vectors.
- What is the key advantage of cosine similarity over Euclidean distance?

Useful for Information Retrieval

- query = 'who wrote wild boys'
- doc1 (D1) = 'Duran Duran sang Wild Boys in 1984.'
- doc2 (D2) = 'Boys don't remain wild forever.'
- doc3 (D3) = 'Who brought wild flowers?'
- doc4 (D4) = 'It was John Krakauer who wrote *In to the wild.'*

Term frequencies

TF	D1	D2	D3	D4	IDF = log(1/DF)
Who	0	0	1	1	$=\log_{10}(1/.5)=.301$
Wrote	0	0	0	1	.602
Wild	1	1	1	1	0
Boys	1	1	0	0	.301

Weights based on document frequency

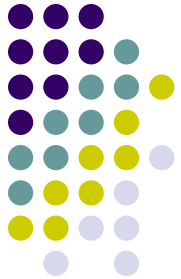
Final TF_IDF weights

TF_IDF	D1	D2	D3	D4	Q
Who	0	0	.301	.301	.301
Wrote	0	0	0	.602	.602
Wild	0	0	0	0	0
Boys	.301	.301	0	0	.301

Similarity	D1	D2	D3	D4
Query	.408	.408	.408	.913*

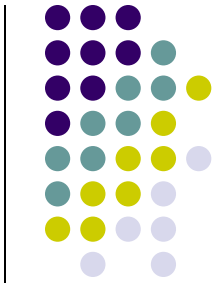
Similarity(Doc1,Query) = (0+0+0+.301^2) / [(.301^2)^.5 * ((.301^2 + .602^2 + .301^2)^.5)]
= .0906 / (.301*.7373) = .408

Application 1: Assessing Relevance and Resonance



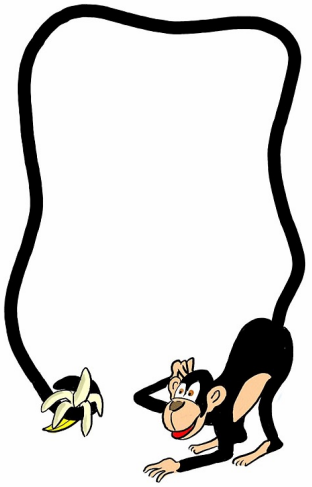
- How do we know if our campaign/message is having an impact on what people are talking about?
  - E.g., political campaign
  - Product or brand campaign
  - How companies describe themselves vs. how employees feel
- A simple approach: How “far” are the social mentions from the message?

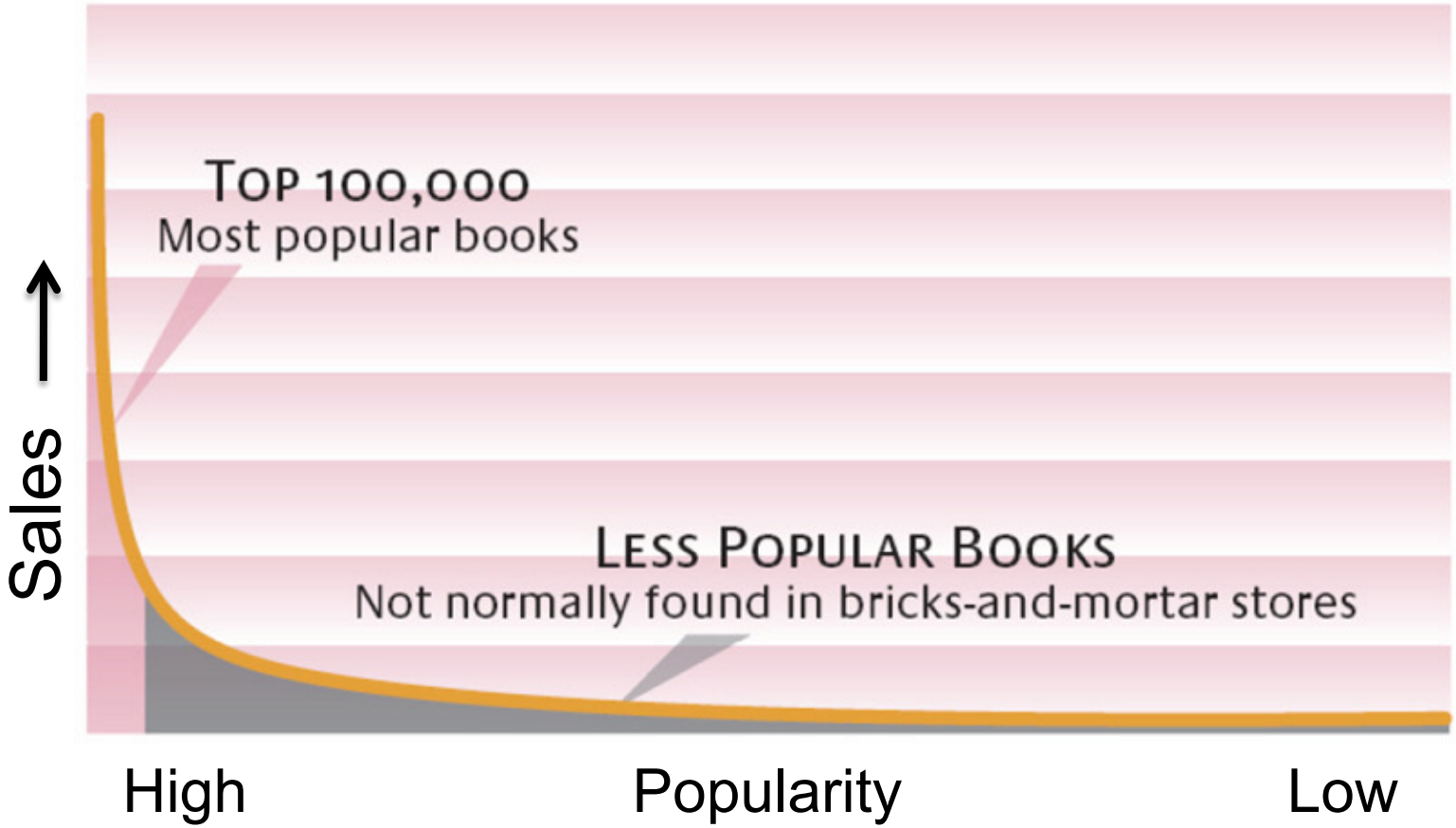
Application 2: Crowdsourced Recommender Systems



- Why bother about recommender systems?
- The long tail of products
- Matching products with customer preferences

Enabling Forces: The Long Tail*

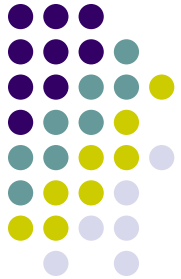




- Customer preferences are diverse
- Endless variety of products available
- Uncertainty in the tail region
- Long tail will not work without recommenders and/or word-of-mouth

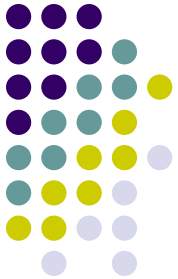
Source: Brynjolfsson, Hu and Smith; *Chris Anderson

Crowdsourced Recommender Systems



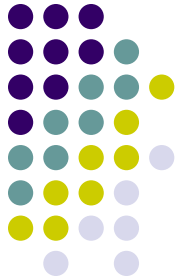
- The case of hotel search
- Web sites (e.g., hotels.com, orbitz.com, Tripadvisor.com) show by one criterion
  - Rating
  - Stars
  - Location
  - Price
- But most customers consider multiple attributes
- How can we provide a better match with customer preferences?

A Lot Better than a Single Criterion



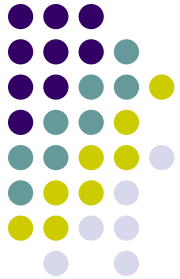
- Obtain user inputs, e.g.,
  - Romantic, ambience, mid-priced, beachfront
  - Business, budget, wi-fi, convenience
- Crowdsource!
- Extract large # conversations about hotels
- Now becomes a document retrieval problem
- E.g., query: Romantic, ambience, mid-price, beachfront
- Calculate cosine similarity between query & mentions
- Rank mentions by cosine similarity
- Perform sentiment analysis to find most positive reviews
- Make recommendations

Additional Data



- Augment text with:
- Vicinity: E.g., Microsoft Virtual Earth Interactive SDK (software development kit)
- Proximity: “near a beach”, “near downtown”, “near public transportation” with user geotagging & automatic classification of satellite images of areas near each hotel (e.g., geonames.org)
- Additional information become extra dimensions for cosine similarity analysis
- Overall results: Superior to those provided by any of the major hotel search sites

Observations on Recommender Systems



- Recommendations may be from the tail region (increases product diversity)
- If you know of a product, a recommender system can find products
  - that are more affordable
  - which people like for the same reasons
- Widespread use of recommenders can shift the demand curve
- For the long tail idea, see https://medium.datadriveninvestor.com/who-is-your-competitor-in-the-era-of-the-long-tail-d0ac24fedde8
- Details of a crowdsourced recommender system: https://towardsdatascience.com/a-recommender-system-based-on-customer-preferences-and-product-reviews-3575992bb61

Craft Beers

The Pursuit of Hoppiness

Anita Bhat
Gihani Dissanayake
Alex Jansen
Kyle Katzen
Siddhant Shah

A “Long Tail” Visibility Problem for Double Sunshine

Double Sunshine IPA | Lawson's Finest Liquids

BA SCORE
4.67/5
2,138 Ratings

BEER STATS

Ranking:	#14
Reviews:	343
Ratings:	2,138
pDev:	7.28%
Bros Score:	0
Wants:	3,276
Gots:	243

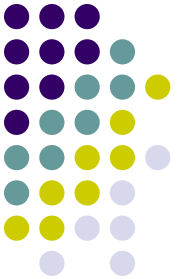
Hopslam Ale | Bell's Brewery, Inc.

BA SCORE
4.46/5
13,068 Ratings

BEER STATS

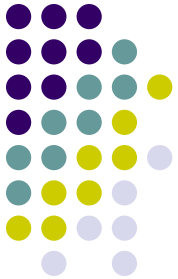
Ranking:	#145
Reviews:	3,532
Ratings:	13,068
pDev:	9.42%
Bros Score:	4.2
Wants:	2,678
Gots:	3,643

Word Embeddings



- A word represented by one number
- What about context?
- Each word being represented by a vector

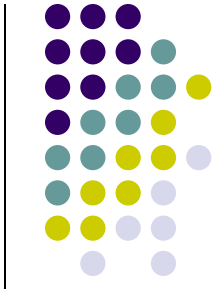
A Dumb Approach ☺



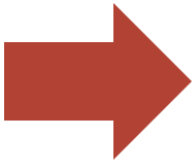
- One-hot encoding
- Also called “1-of-N” encoding
- Embedding space has the same number of dimensions as the number of words in the vocabulary
- Each embedding is basically made up of zeros, with a “1” in the corresponding dimension for the word

Source: https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/

One-hot Encoding for a 9-word Vocabulary



Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch



	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Each word gets a 1x9 vector representation

Issues with this approach?

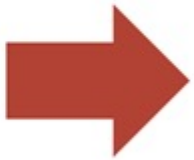
Source: https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/

Reducing Dimensionality: Custom Embeddings

Goals:

- Reduce dimensionality
- Similar words → similar vectors

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch

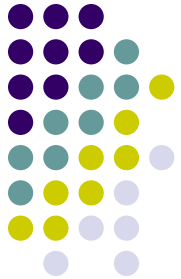


	Feminine	Youth	Royalty
Man			
Woman			
Boy			
Girl			
Prince			
Princess			
Queen			
King			
Royalty			

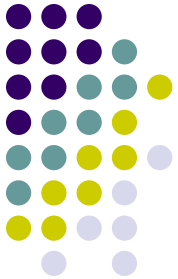
Each word will have 1x3 vector representation

Source: https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/

Advantages of the New Encoding?



Larger Vocabularies



- Creating n-dimensional vectors from large corpus
- Manual assignment of vectors not possible
- Word embeddings should have hundreds of dimensions
- Vector values can be assigned in a variety of ways (including prediction, which is most useful)
- Algorithms take large bodies of text and create embeddings
- Word2Vec (Mikolov et al., Google), GloVe (Stanford) & fastText (Facebook)

Source: https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/

Automating Word Vector Generation

John Rupert Firth

"You shall know a word by the company it keeps" (Firth, 1957)

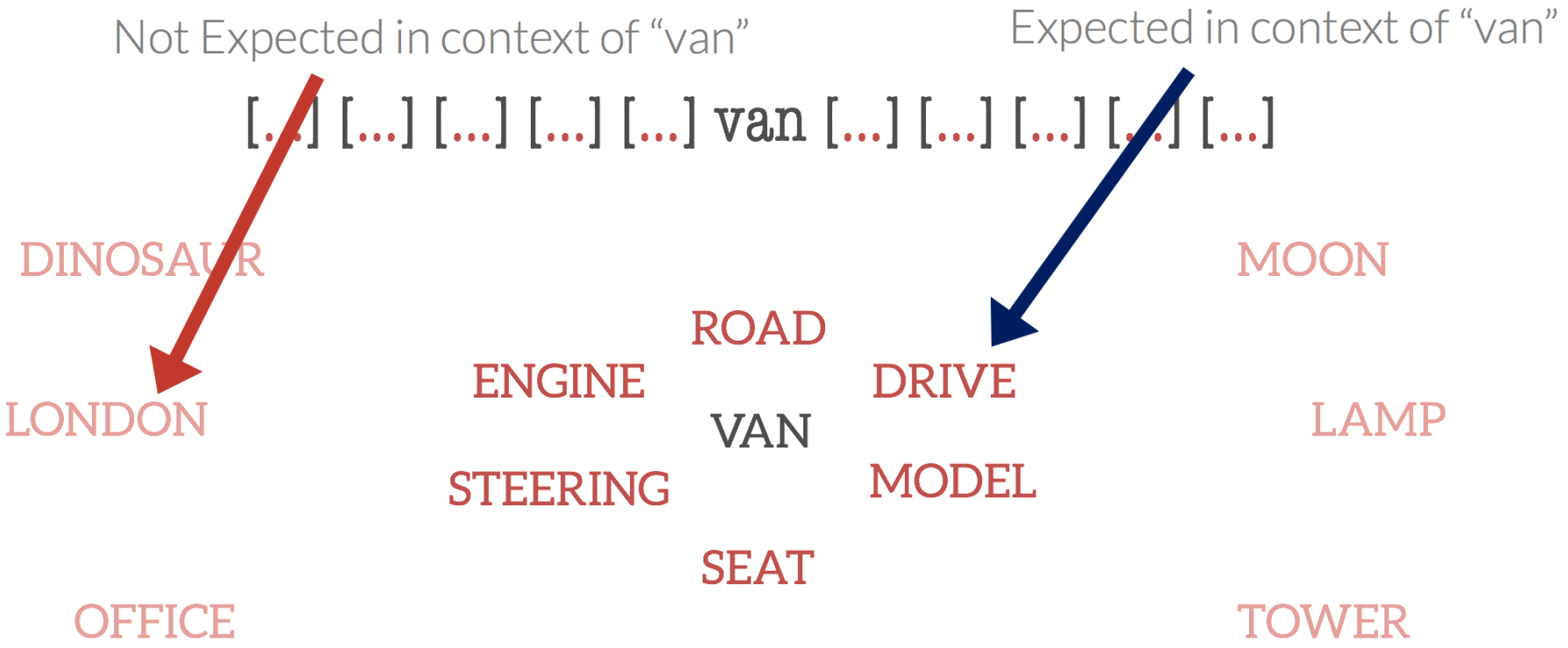
[drove] [my] [high] [speed] vehicle [down] [the] [road] [today]

Center word

Context

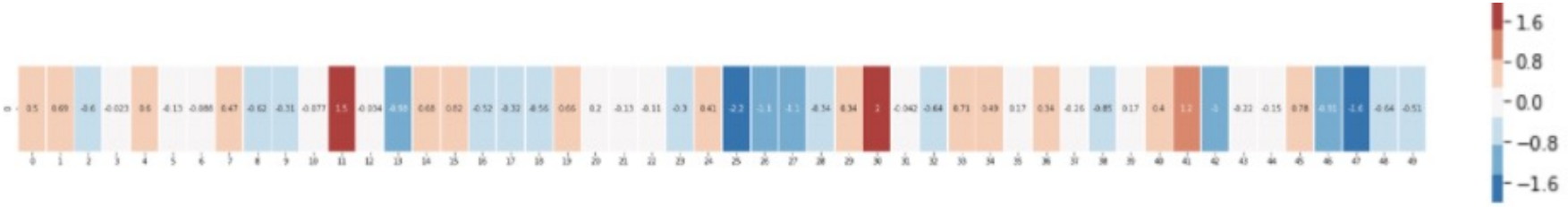
Context

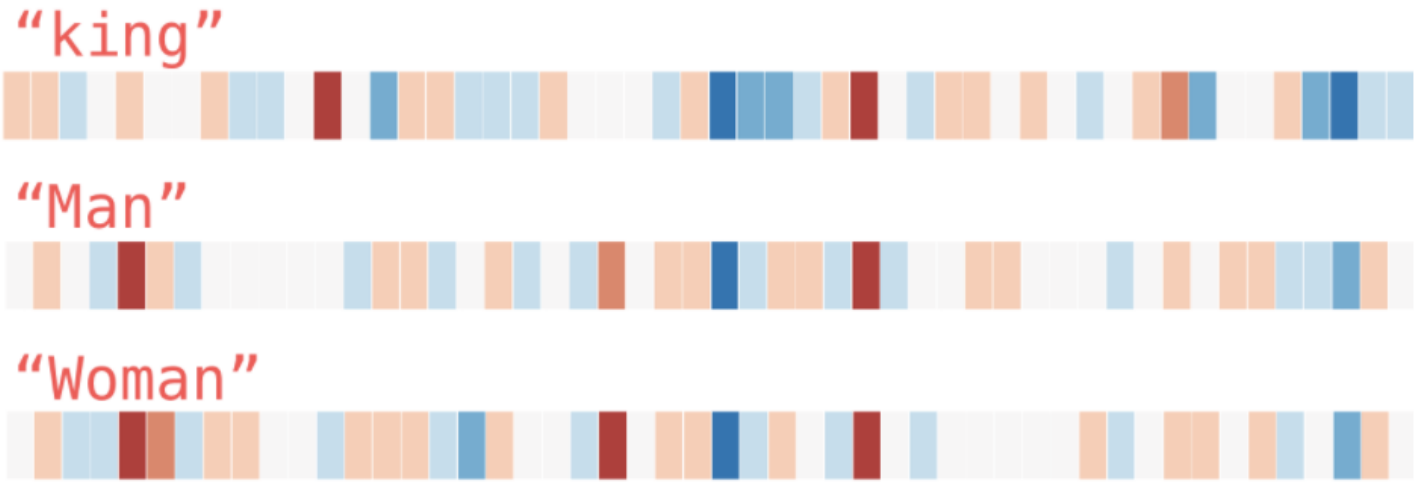
Given a Center Word, Find Context Words



Pre-trained Word Embeddings

- The 50-dimensional Glove embeddings for "**king**"
- [0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961 , -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 , -0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 , -1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042]
- The numbers are weights in the neural network used for training (details coming up!)

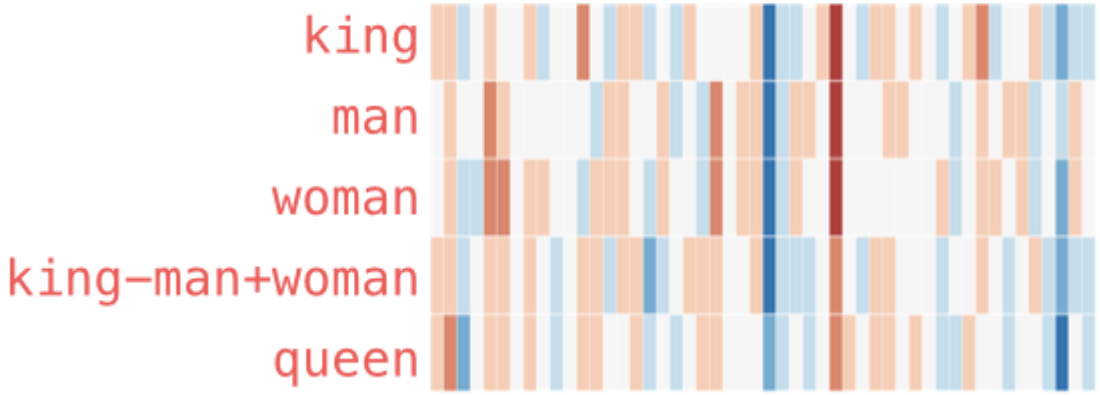




Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

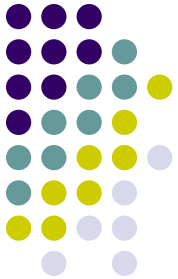
Vector Addition and Subtraction: The Case of Analogies

king − man + woman ~= queen



Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

Language Models



- Fill in the blank
- She was driving a ____
- Would your answer change for
- She was driving a ____ red Ferrari
- Words to both left and right of a center word have informational value
- Window of size 5 (just an example)

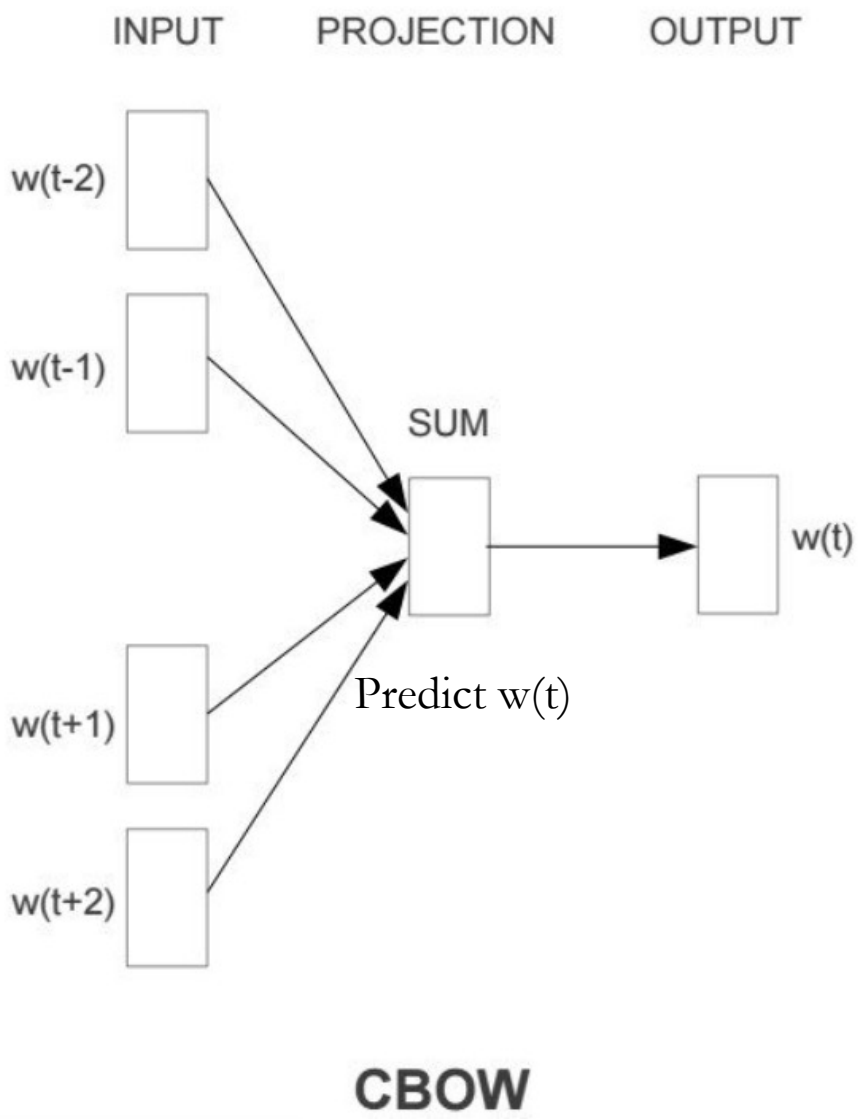
Driving	a	cool	red	Ferrari

Input 1	Input 2	Input 3	Input 4	Output (center word)
Driving	a	red	Ferrari	cool

Continuous Bag of Words (CBOW): Given the context words, predict the center word

Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

Continuous Bag of Words (CBOW)



Another Model: Skip-gram

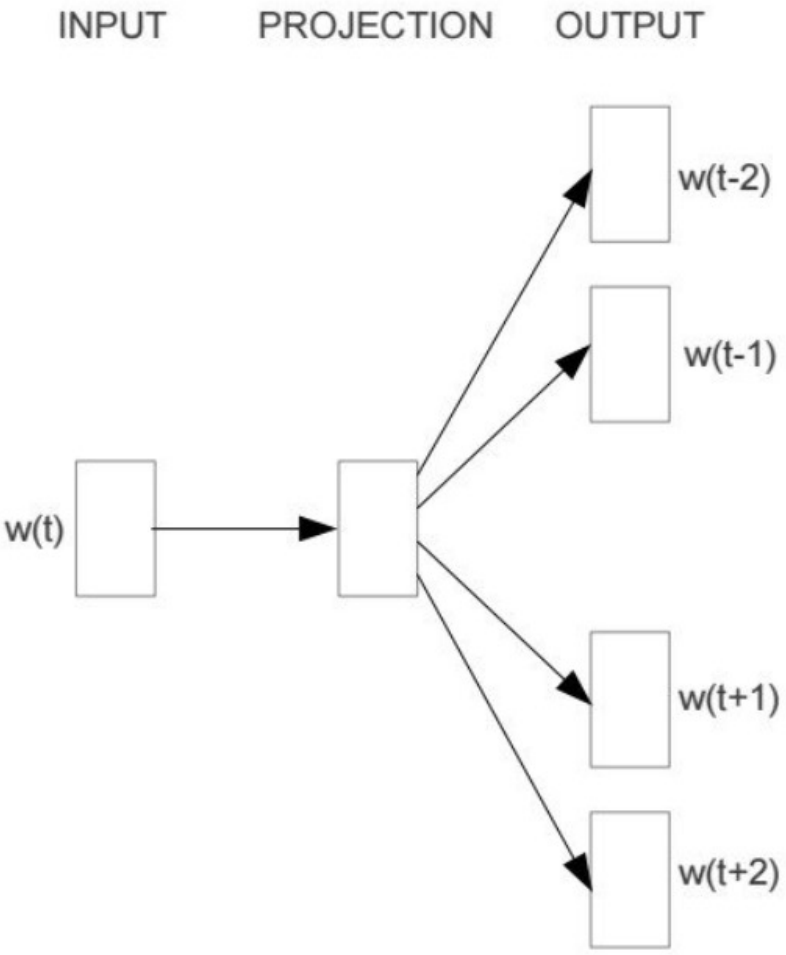
Driving	a	cool	red	Ferrari

Input	Output (context words)
cool	Driving
cool	a
cool	red
cool	Ferrari

Two words to the left and right of a center word (just an example)
Skip-gram: Given the center word, predict the context words

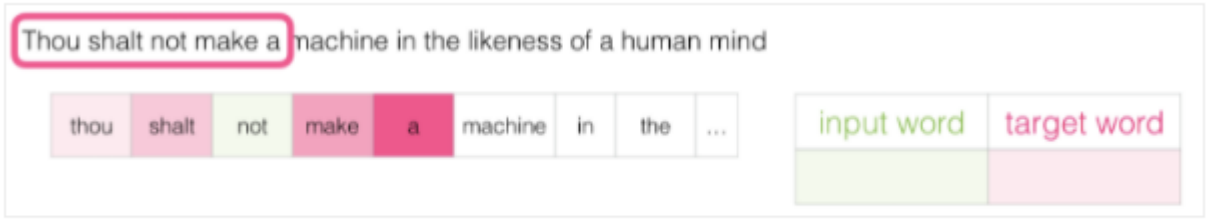
Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

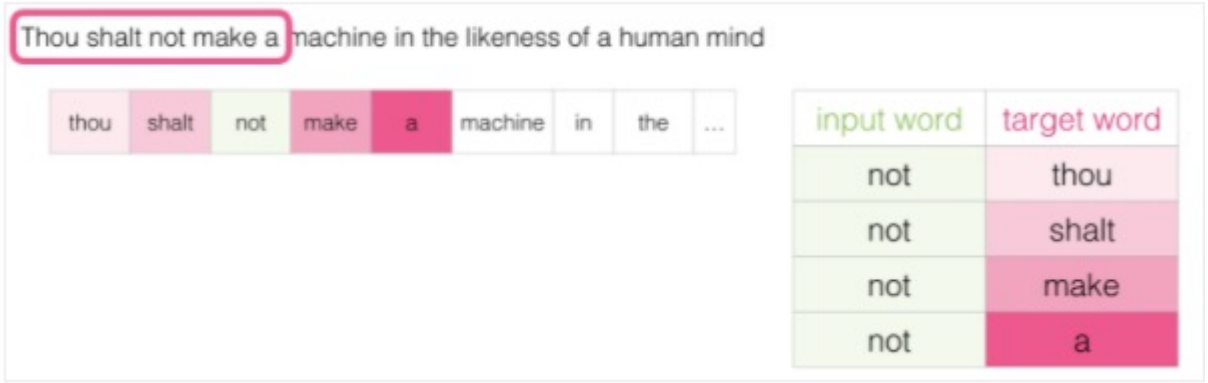
The Skip-gram Model



Skip-gram

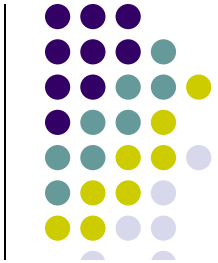
The Sliding Window





Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

After Sliding Over Several Words



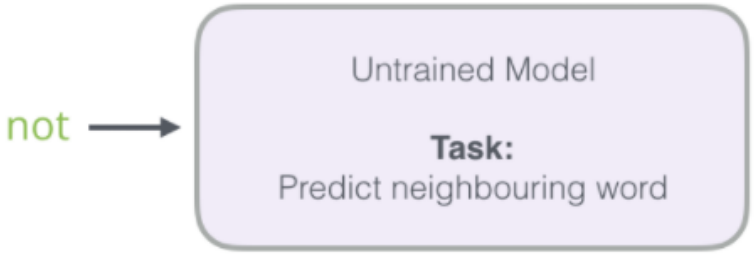
Thou shalt not make a machine in the likeness of a human mind

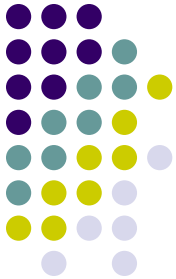
thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	…

input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness

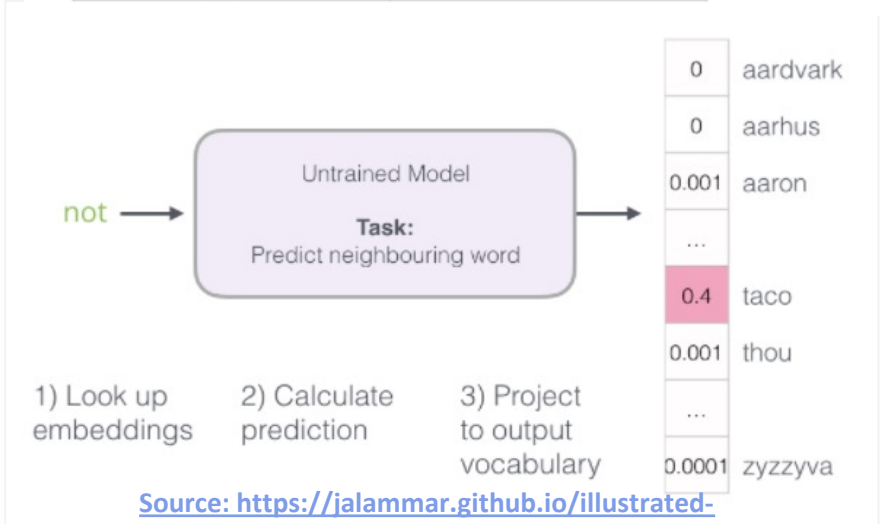
Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness



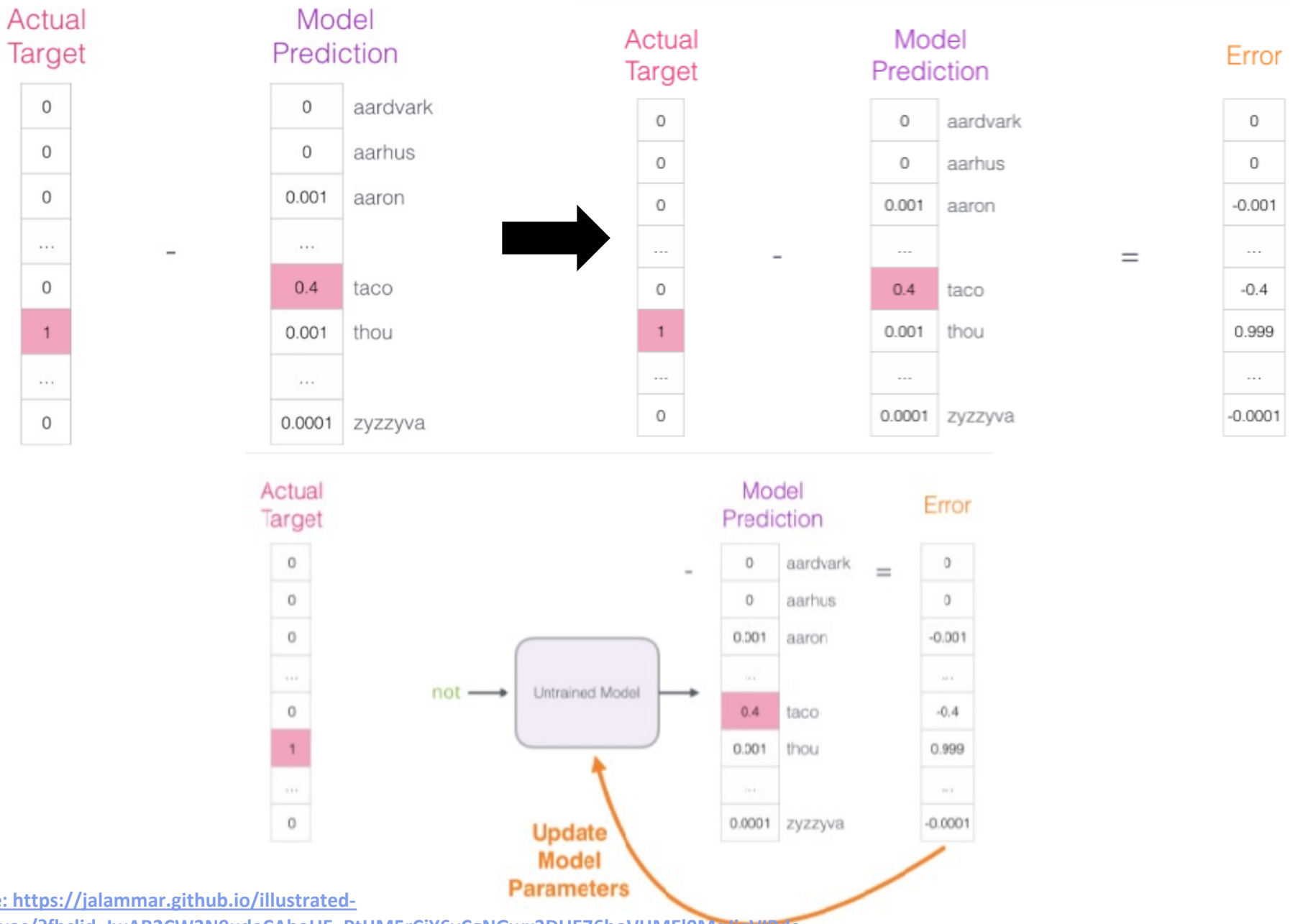


Start the Training



Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

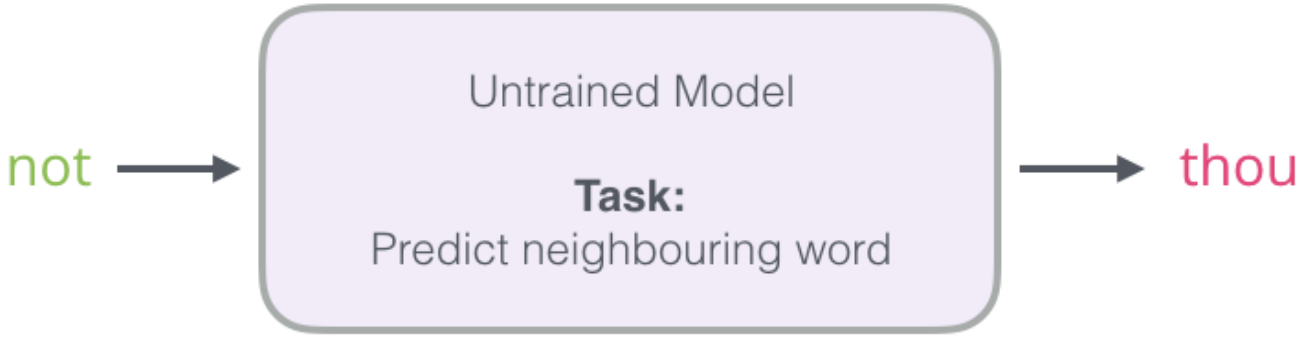
Updating Model Parameters



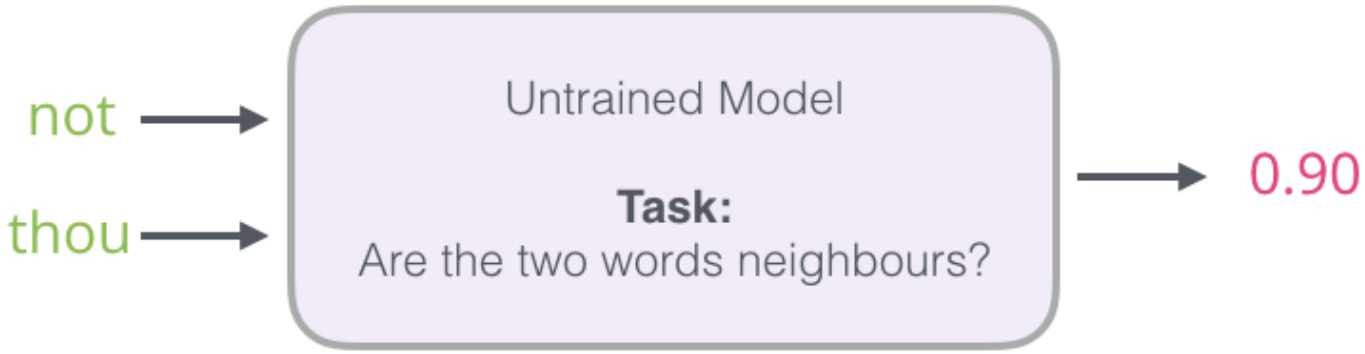
Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCiY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

Increase the Efficiency of the Model

From



To



input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine

input word	output word	target
not	thou	1
not	shalt	1
not	make	1
not	a	1
make	shalt	1
make	not	1
make	a	1
make	machine	1

Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCiY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

Negative Sampling

The problem

input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine

input word	output word	target
not	thou	1
not	shalt	1
not	make	1
not	a	1
make	shalt	1
make	not	1
make	a	1
make	machine	1

The solution

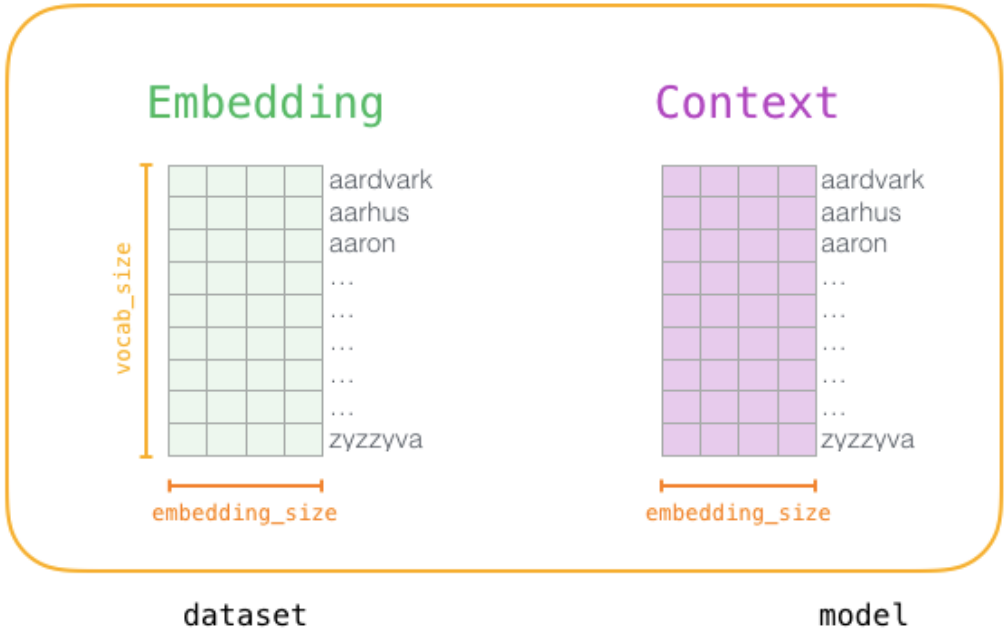
input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	make	1

Pick randomly from vocabulary (random sampling)

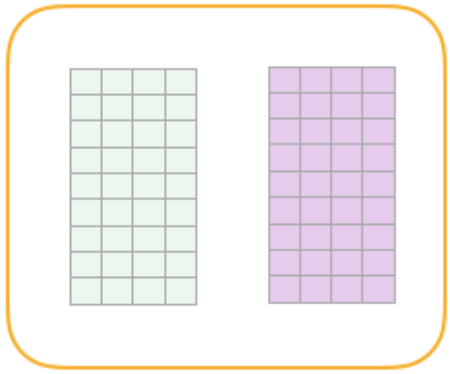
Word	Count	Probability
aardvark		
aarhus		
aaron		
taco		
thou		
zyzzyva		

Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

All Set for Training

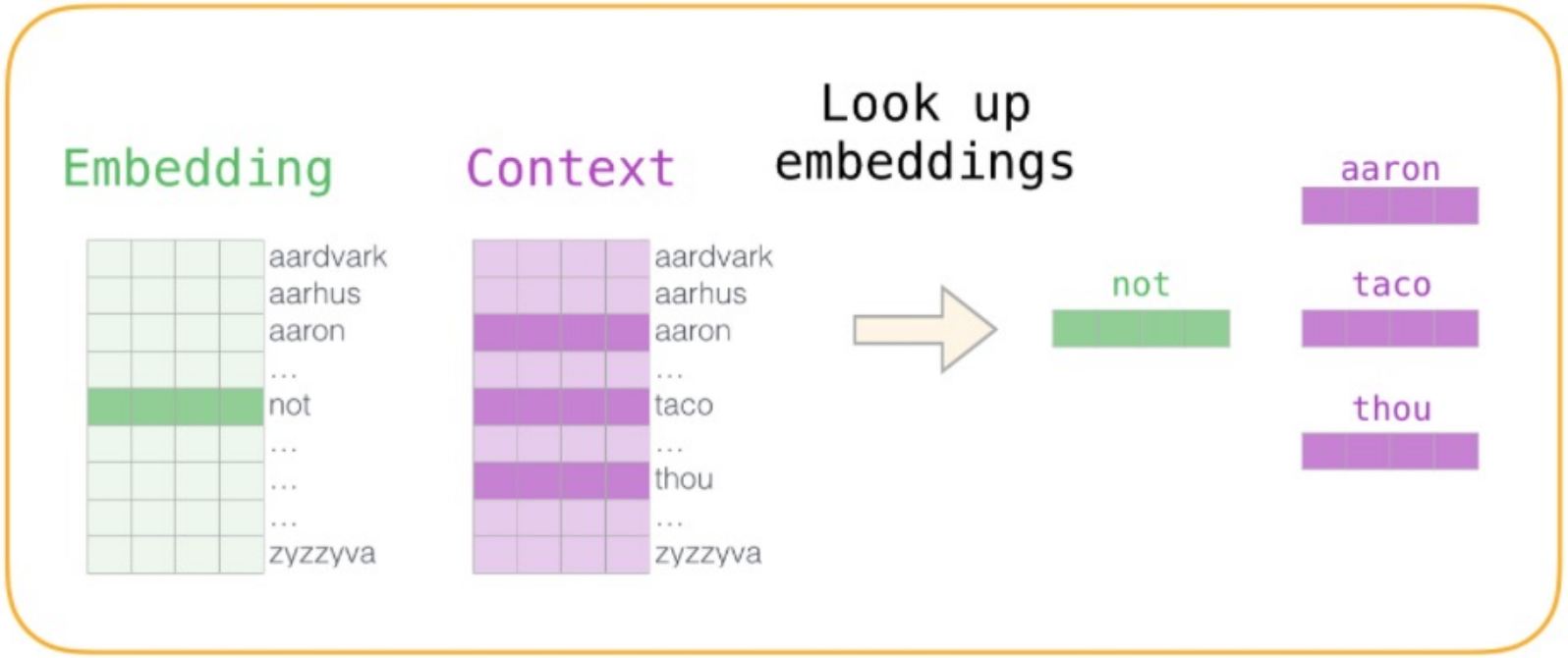


input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	mango	0
not	finglonger	0
not	make	1
not	plumbus	0
…	…	…

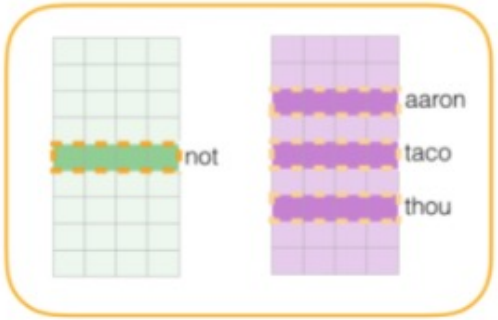


Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

The Training Process



input word	output word	target	input • output	sigmoid()	Error
not	thou	1	0.2	0.55	0.45
not	aaron	0	-1.11	0.25	-0.25
not	taco	0	0.74	0.68	-0.68

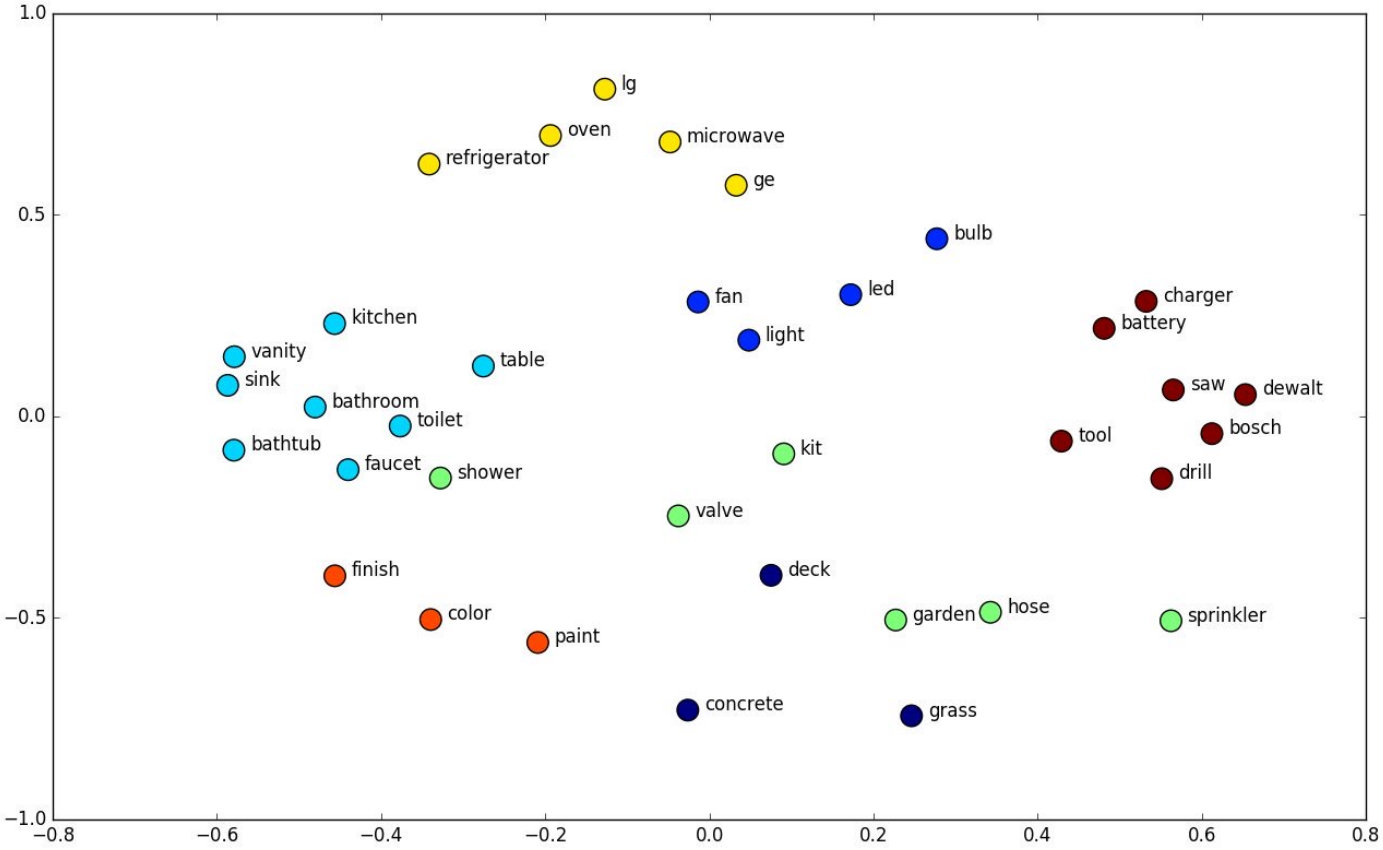


Update Model Parameters

Source: https://jalammar.github.io/illustrated-word2vec/?fbclid=IwAR2CW3N9udeCAboUE_PtHM5rCjY6vCgNGwx3DUE76boVHMEl0MqJigVIBdc

Word Similarities

- In trained word vectors, similar words will be close to each other
- Similarity can be calculated by Euclidean distance or cosine similarity of word vectors



Discovery of “New” Words

1. Nearest neighbors

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *frog*:

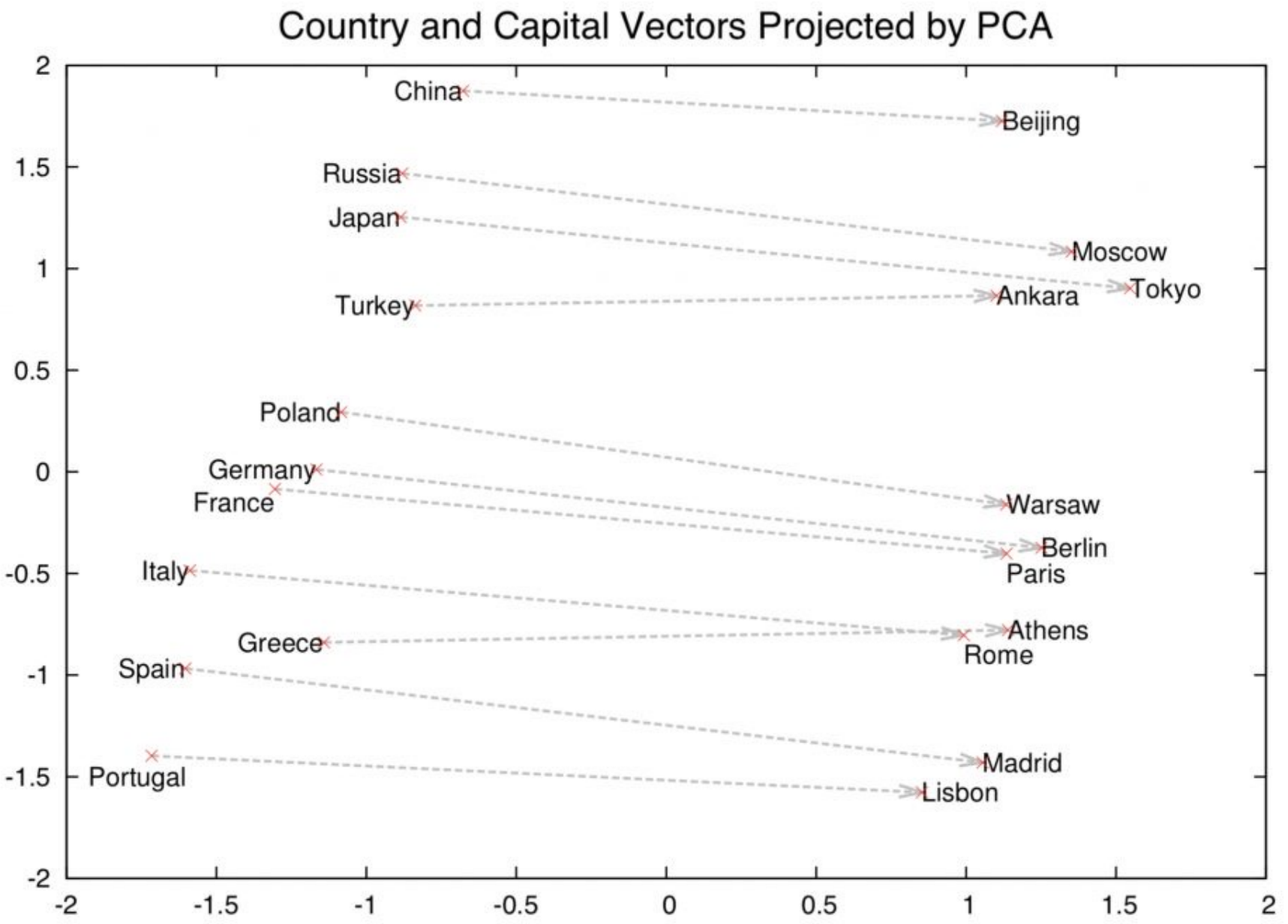
0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

3. litoria

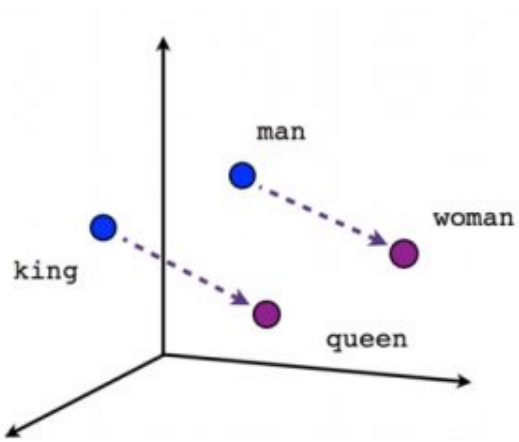
4. leptodactylidae

5. rana

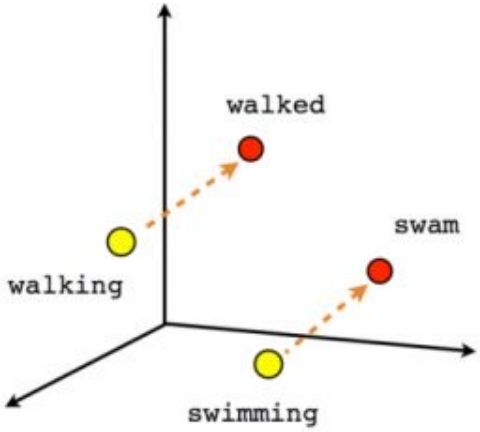
7. eleutherodactylus


Country and Capital Vectors Projected by PCA
China
Beijing
Russia
Moscow
Japan
Tokyo
Turkey
Ankara
Poland
Warsaw
Germany
Berlin
France
Paris
Italy
Athens
Rome
Greece
Spain
Madrid
Portugal
Lisbon
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

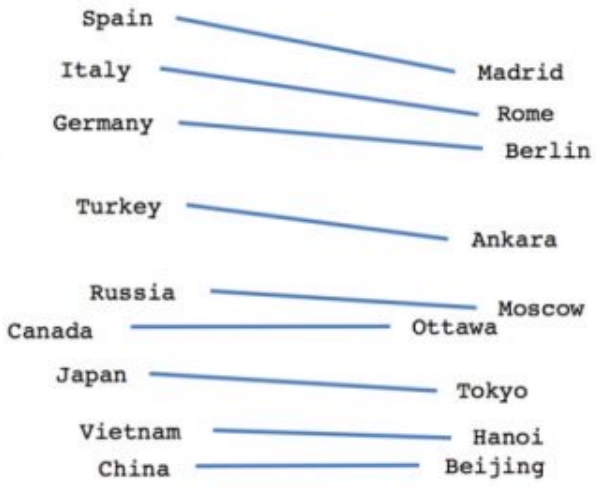
Discovering Relationships from Corpus



Male-Female



Verb tense



Country-Capital

UNSTRUCTURED DATA ANALYTICS

Sentiment Analysis

MSITM Fall 2022
26th September

Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor of Business
Distinguished Fellow, INFORMS Information Systems Society
Stevens Piper Foundation Professor
University of Texas Distinguished Teaching Professor
McCombs School of Business, University of Texas at Austin
Email: **aniteshb@gmail.com**

Incorporating Sentiments & Mentions in the Stock Returns Model

- Do message sentiments improve prediction of returns?
- How about news items?
- Training data for stock sentiments

appleup • 16 minutes ago

more revenue coming.....by the billions.....Apple style !

Sentiment: Strong Buy

bistec_98 • 16 hours ago

WOW! Look at the DROP! NICE

Nice Correction.

Sentiment: Strong Sell

Tim Cook and the Apple management team has accomplished so much with new products and new market developments over the years.....and still has the Apple brand image being known all around the world as the best quality products available on the market. The management team has been so creative and... More

Sentiment: Strong Buy

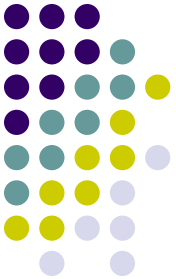
bernsteinjoshi • Aug 27, 2015 10:54 AM 1 1 Reply

NO One with a Brain would touch a Scam which has a Book Value of laughable $25 and $18 Billion in Debt

especially when the Company is unprofitable .The Market Cap of astronomical $240 Billion is far away from reality .Next point is How will the Company reach profitability when the World Economy is collapsing ? They didnt reach it even when the Economy was stable .

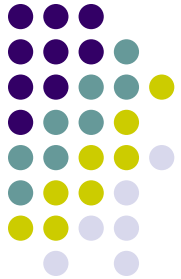
Sentiment: Strong Sell

Predicting Stock Returns (the Trillion $ Question): Do Sentiments Matter?



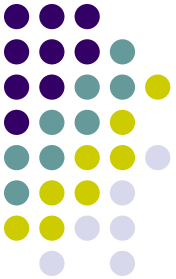
- Capital Asset Pricing Model (CAPM)
- Fama-French factors
  - Return on an asset
  - $r = R_f + \beta.(K_m - R_f) + b_s.SMB + b_v.HML + \alpha$
  - R_f: risk free rate, K_m: Return of market portfolio
  - *SMB:* Small (market cap) minus big
  - *HML:* High (book-to-mkt) minus low
- Addition of momentum (Up minus Down) factor

Unsupervised Sentiment Analysis



- Many approaches, but all use reference positive & negative words
- Lexicons with either binary tags or valence (intensity of sentiment)
- Can incorporate other features – e.g., bigrams, sentiment shifters, etc.
- Parts-of-speech (POS) tagging
  - POS: Noun, verb, adjective, adverb, pronoun, preposition, conjunction & interjection
  - Extract adjectives
  - Isolated adjectives may not indicate true opinion orientation
  - Extract two consecutive words if their POS conform to a pattern
- Studied heavily in linguistics

Extract Two-Word Phrases Using Parts-of-Speech Tagging

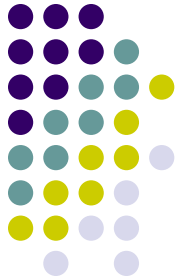


Rule	First word	Second Word	Third word (not extracted)
1.	JJ (Adjective)	NN or NNS	Anything
2.	RB (Adverb), RBR or RBS	JJ	Not NN or NNS
3.	JJ	JJ	Not NN or NNS
4.	NN (Noun) or NNS	JJ	Not NN or NNS
5.	RB, RBR or RBS	VB (verb), VBD, VBN or VBG	Anything

E.g., “This business runs like a virtual monopoly.”

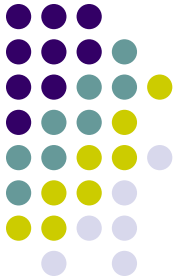
Tagging tool: http://cogcomp.cs.illinois.edu/demo/pos/?id=4

Mining the Sentiment



- Associate extracted phrases with
  - positive reference words like “excellent”, “great”, etc.
  - negative reference words like “poor”, “terrible”, etc.
  - See which association is dominant

Unsupervised Learning Using POS



- Point-wise mutual information (PMI) of two phrases (or words) = amount of information we get about the presence of one phrase given the presence of the other.
- *p*(phrase 1 **and** phrase 2) / [*p*(phrase 1) * *p*(phrase 2)] shows how statistically dependent the phrases are
  - where *p* is the probability
  - E.g., *p*("virtual monopoly" and "awesome") / [*p*("virtual monopoly") * *p*("awesome")]
- Taking log of the above gives us the "quantity" of information we get about one phrase in the presence of the other
- PMI = $\log_2$ [*p*(phrase 1 and phrase 2) / *p*(phrase 1) * *p*(phrase 2)]

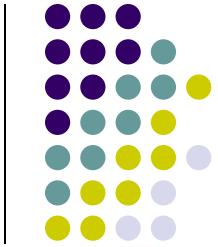
Calculating Semantic Orientation

- Opinion orientation (a.k.a. semantic orientation) = PMI("virtual monopoly", "awesome) – PMI("virtual monopoly", "terrible") = $\log_2$ of the ratio below:

cases of "virtual monopoly" AND "awesome" * # cases of "terrible"

#cases of "virtual monopoly" AND "terrible" * # cases of "awesome"

Applying POS Bigrams for Opinion Mining



Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
little difference	JJ NN	-1.615
clever tricks	JJ NNS	-0.040
programs such	NNS JJ	0.117
possible moment	JJ NN	-0.668
unethical practices	JJ NNS	-8.484
low funds	JJ NNS	-6.843
old man	JJ NN	-2.566
other problems	JJ NNS	-2.748
probably wondering	RB VBG	-1.830
virtual monopoly	JJ NN	-2.050
other bank	JJ NN	-0.850
extra day	JJ NN	-0.286
direct deposits	JJ NNS	5.771
online web	JJ NN	1.936
cool thing	JJ NN	0.395
very handy	RB JJ	1.349
lesser evil	RBR JJ	-2.288
Average Semantic Orientation		-1.218

Processing of a review where author summarized as “not recommended”

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
small part	JJ NN	0.053
online service	JJ NN	2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	1.288
well other	RB JJ	0.237
inconveniently located	RB VBN	-1.541
other bank	JJ NN	-0.850
true service	JJ NN	-0.732
Average Semantic Orientation		0.322

Processing of a review where author summarized as “recommended”

Valence Aware Dictionary for sEntiment Reasoning (VADER)

A popular tool for unsupervised sentiment analysis
Available in Python

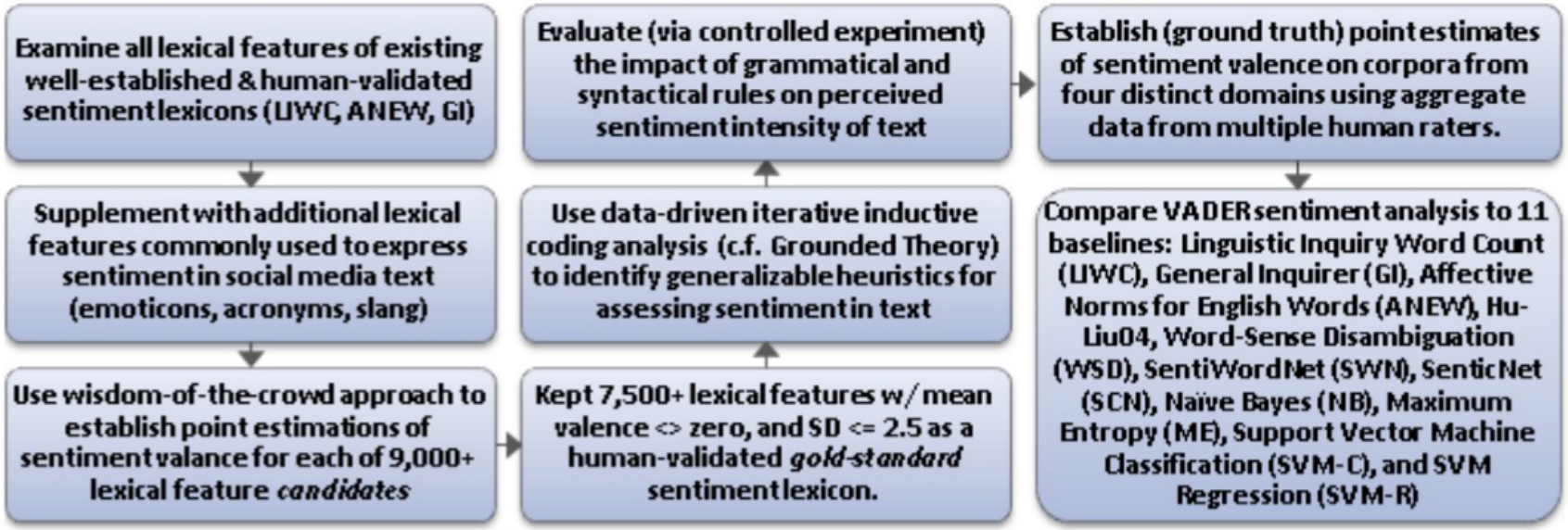
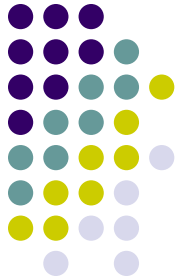


Figure 1: Methods and process approach overview.

Source: Hutto & Gilbert 2014

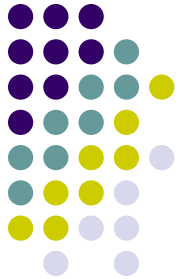
Adds Human Heuristics for Expressing Sentiments



- Punctuation, e.g., ! increases the sentiment intensity
- Capitalization, e.g., ALL-CAPS
- Degree modifiers, e.g., “The service here is extremely good”
- Sentiment shifters, e.g., “The food here is great, but the service is horrible”
- Trigram preceding a sentiment-laden lexical feature: Catch nearly 90% of cases where negation flips the polarity of the text.
- E.g., “The food here isn’t really all that great”.

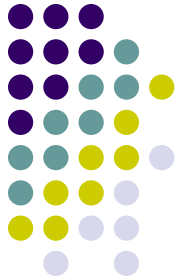
Source: Hutto & Gilbert 2014

Multiple Products, Multiple Attributes



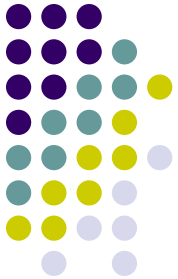
- *"The lobby of the Coastal Delight was cool, but the room was nothing to write home about; the food was good, but the location was too far away from public transportation. In hindsight, although the Tuscany Grand was very pricey, its awesome location and high-end ambience would have been great."*
- Assumption: People express emotions in close proximity to the mentions of entities and/or attributes.
- Parse and extract phrases that are "relevant"
- E.g., *Dell* & *warranty*, *hotel* & *lobby*, etc.
- Then pass through a sentiment analyzer like VADER

Supervised Sentiment Analysis



- Same as classification or prediction with text
- Start with a collection of documents whose sentiments are known (i.e., they have been labeled manually or otherwise)
- Many approaches are possible – from the naïve to the sophisticated
- Unigrams or “bag-of-words” for starters – single words, assumed to occur independently of each other in a document

Simplest Approach to Sentiment Analysis?



- "It's rather like a lifetime special – pleasant, sweet and forgettable."

Positive: 506
Negative: 507
Goodness score:
Badness score:

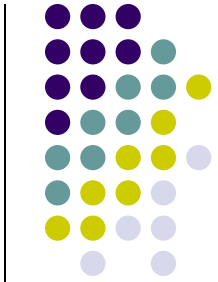
Positive: 46
Negative: 22
Goodness score:
Badness score:

Positive: 10
Negative: 14
Goodness score:
Badness score:

Positive reviews in training data: 15 occurrences
Negative reviews: 6
Goodness score:
Badness score:

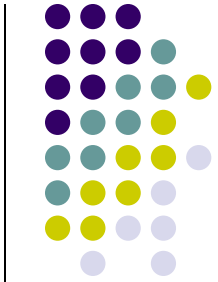
Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

Calculations



Words	#Positive	#Negative	Goodness	Badness
it's	506	507	.5	.5
rather	42	63	.4	.6
like	242	396	.61	.39
a	3446	3112	.53	.47
lifetime	3	5	.38	.62
special	29	40	.42	.58
sweet	15	6	.71	.29
pleasant	46	22	.68	.32
and	3198	2371	.57	.43
forgettable	10	14	.42	.58
SCORE			**5.22**	**4.8**

Top-10 Words



Positive	Negative
riveting	unfunny
gem	badly
engrossing	poorly
vividly	flat
wonderfully	bore
polished	pointless
lively	offensive
heartwarming	plodding
startling	product
spare	disguise

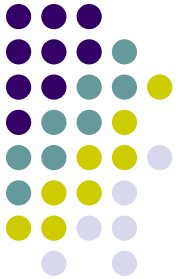
Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

Tough to Detect

- This movie makes Catwoman look like a great movie
- A terrible movie that some people will nevertheless find very moving
- Well made but mush-hearted
- Your children will be occupied for 72 minutes

Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

K-Nearest Neighbors

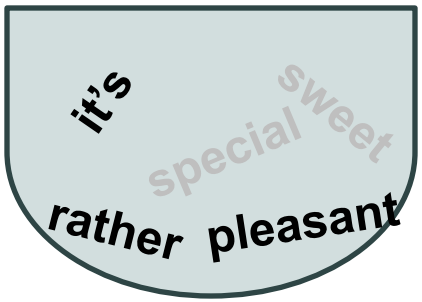


- Instead of comparing individual words in the training set, consider entire reviews.
- Find “nearest” reviews
- Example
  - Classify as positive or negative:
    - “It’s rather like a lifetime special – pleasant, sweet and forgettable”
  - Find “most similar” review(s) in the training set
    - “I liked this movie, made for a pleasant evening” (Class = positive)
    - “This movie was such a bore” (Class = negative)

Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

The Difference

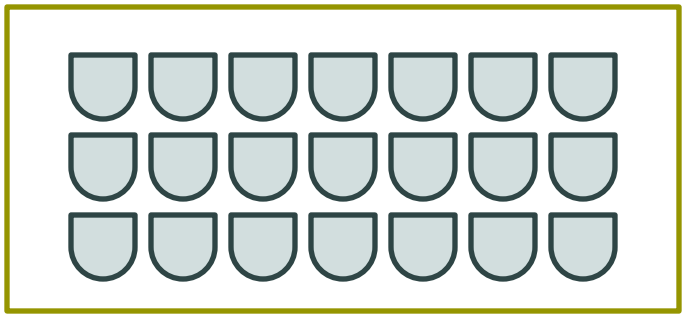
BEFORE



NOW

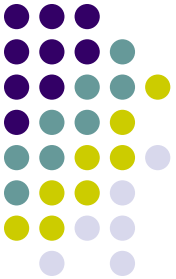
compare

TRAINING SET
BAGS



Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

Compare with Reviews in Training Set



Task: To determine the sentiment of
"It is a surprisingly funny movie!"

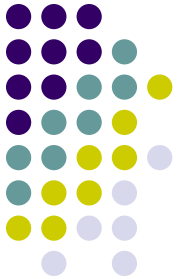
[surprisingly, funny, movie]

A review in the training data labeled as Positive:
"funny movie, I recommend it."

[funny, movie, recommend]

Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

How do we Quantify the Distance?



[surprisingly, funny, movie]

"funny" occurs in, say, **10%** of all reviews

"movie" occurs in, say, **70%** of all reviews

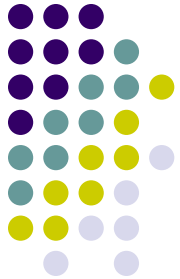
[funny, movie, recommend]

Need to know frequencies of matching words in the training data

Similarity score = 1/0.1 + 1/0.7 = 10 + 1.43
= 11.43

Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

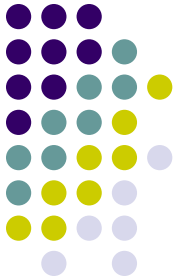
Importance of Words

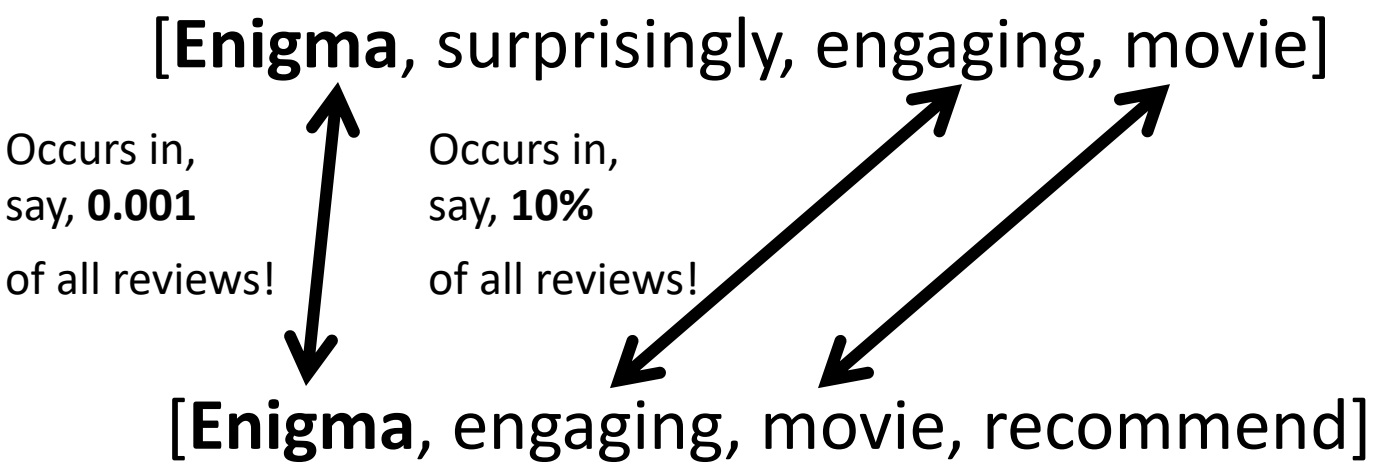


- Classify
  - “Enigma is a surprisingly engaging movie”
- One review from training set
  - “Enigma is a movie that engrosses you, I recommend it.”
  - More info: “Enigma” occurs .001 (1 in 1000 times)
  - “Engage/engross” occurs .3 (30% of all training set reviews)
  - “Movie” occurs .7 (70%)
- How do we create a similarity score?

Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

Add a Little Twist ….





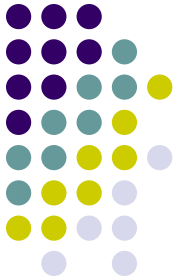
Similarity score = 1/0.001 + 1/0.1 + 1/0.7

= 1000 + 10 + 1.43

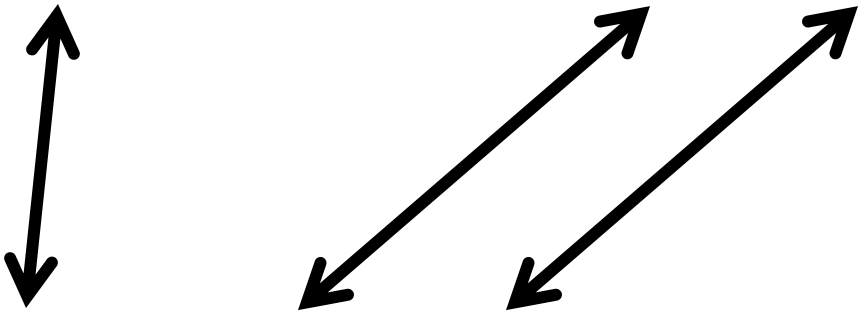
= 1011.43

Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

Not all Words are Created Equal



[Enigma, surprisingly, funny, movie]



[Enigma, funny, movie, recommend]

SCORE= log(1/0.001) + log(1/0.1) + log(1/0.7)
= log(1000) + log(10) + log(1.43)
= 3 + 1 + 0.15 = 4.15

Source: https://www.youtube.com/watch?v=0JsHvXmU0dA

Unstructured Data Analytics

Accessing Data
NLP Fundamentals

MSITM, Fall 2022, Session 3, September 12

Dr. Anitesh Barua

David Bruton Jr. Centennial Chair Professor in Business

Distinguished Fellow, INFORMS Information Systems Society

University of Texas Distinguished Teaching Professor

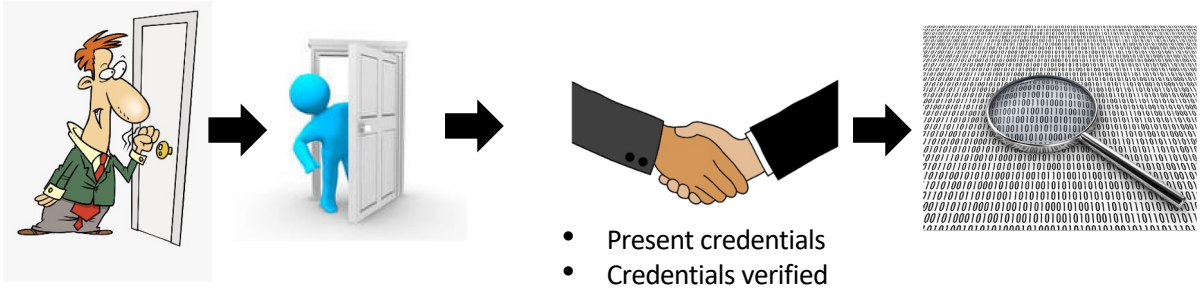
Associate Director, Center for Research in e-Commerce

McCombs School of Business, University of Texas at Austin

Email: **aniteshb@gmail.com**

Accessing Data from Websites

If the Website has an API: The Polite Way!

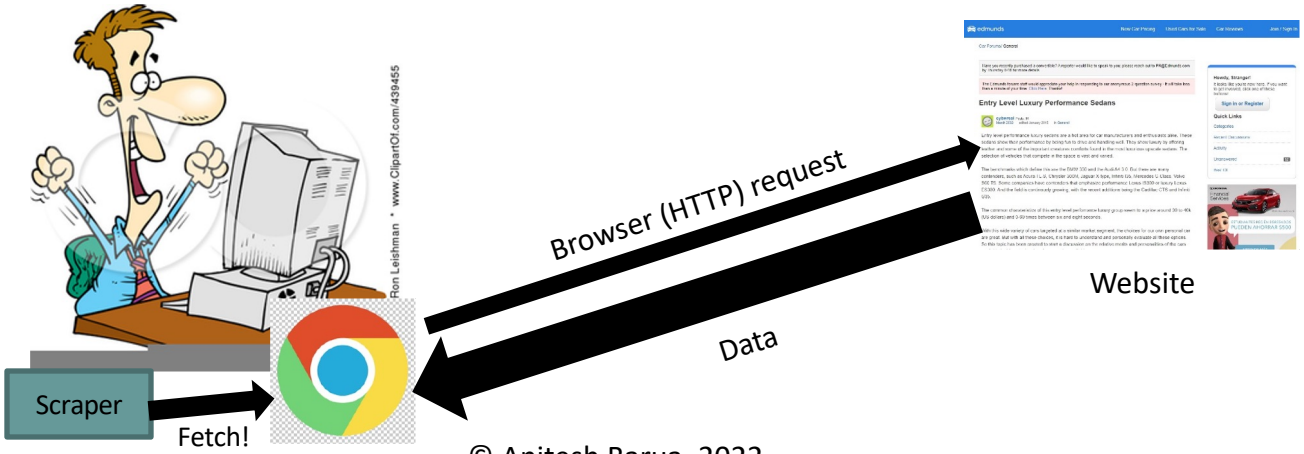


- Present credentials
- Credentials verified
- Request data

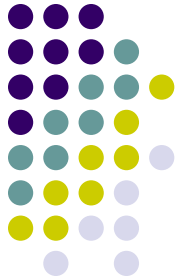
- E.g., Twitter API
  - Limitations: Only 3-4 days old data, limit on # tweets (~3000 per search)
  - Advantages: Can get detailed data (e.g., user location)

No API? We'll Take it Anyway ☺

- Two methods
- Send a *bot* to the site to 'crawl' the data
- E.g., Scrapy and beautifulsoup in Python
  - Fast
  - Can fetch a lot of data (esp. Scrapy)
  - But easy to detect by the site
  - Can easily get blocked
- Hiding behind your browser (e.g., with Selenium)

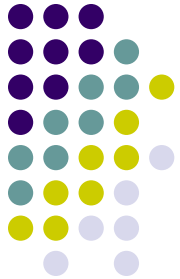


Multiple Ways to Scrape



- Many scraping tools available (don’t require any coding)
- E.g., WebScraper, Octoparse, etc.
- But can’t be used for assignment 1 (OK to use for others + final project)

Dealing with Text



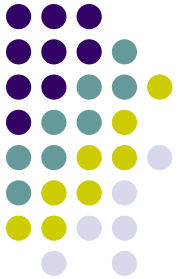
- From a stream of characters to *tokens*
- Sentence and word-level tokenization with nltk
  - E.g., 'Hello Mr. Smith, how are you doing today?', 'The weather is great, and Python is awesome.', 'The sky is pinkish-blue.', "You shouldn't eat so many cookies.“

```
from nltk.tokenize import sent_tokenize, word_tokenize
TEXT = "Hello Mr. Smith, how are you doing today? The weather is great, and Python is awesome. The sky is pinkish-blue. You shouldn't eat so many cookies."
print(sent_tokenize(TEXT))
print(word_tokenize(TEXT))
```

* https://pythonprogramming.net/tokenizing-words-sentences-nltk-tutorial/ for tokenization

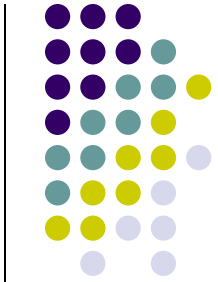
Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Part-of-speech



https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Tagging Tokens

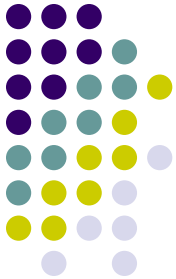


- Parts-of-speech (POS) tagging

```
import nltk
from nltk.tokenize import word_tokenize
text=word_tokenize("And now for something completely different")
print(nltk.pos_tag(text))
```

* Need to install numpy for pos_tag: http://sourceforge.net/projects/numpy/files/

Unigrams Vs. Bi- or Trigrams

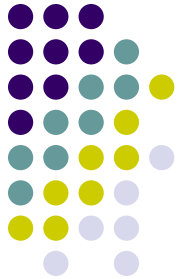


- Sometimes bigrams may be more meaningful than unigrams

```
import nltk
from nltk import word_tokenize
from nltk.util import ngrams
from collections import Counter
text = "This camera produces awesome pictures."
token = nltk.word_tokenize(text)
bigrams = ngrams(token,2)
print(Counter(bigrams))
trigrams = ngrams(token,3)
print(Counter(trigrams))
```

- What about POS bigrams? When would they be useful?

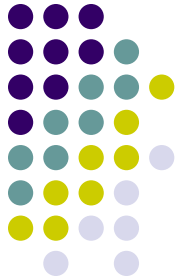
Stemming & Lemmatization



- Reducing the size of the vocabulary in a doc collection
- Different forms (inflections) of a word (e.g., *organize*, *organizes*, *organized* and *organizing*)
- Families of derivationally related words with similar meanings:
  E.g., *democratic*, *democratization*.*
- Stemming: *"The process for reducing inflected (or sometimes derived) words to their stem…"* +
  - A heuristic process that chops off the ends of words.
  - Operates on single words without knowledge of context
- Check out http://textanalysisonline.com/nltk-porter-stemmer

*Stanford NLP Group, +Wikipedia

Lemmatization



- *"The process of grouping together the different inflected forms of a word so they can be analyzed as a single item."*
- Lemma = root or base word
  - E.g., 'walk', 'walked', 'walks', 'walking' -> walk
- More complex than stemming: Requires understanding of context, determining the part of speech, etc.
- E.g., lemmatize 'is' or 'are' as verbs
- Check out http://textanalysisonline.com/spacy-word-lemmatize

Linking Occurrence of Words to Zipf's Law

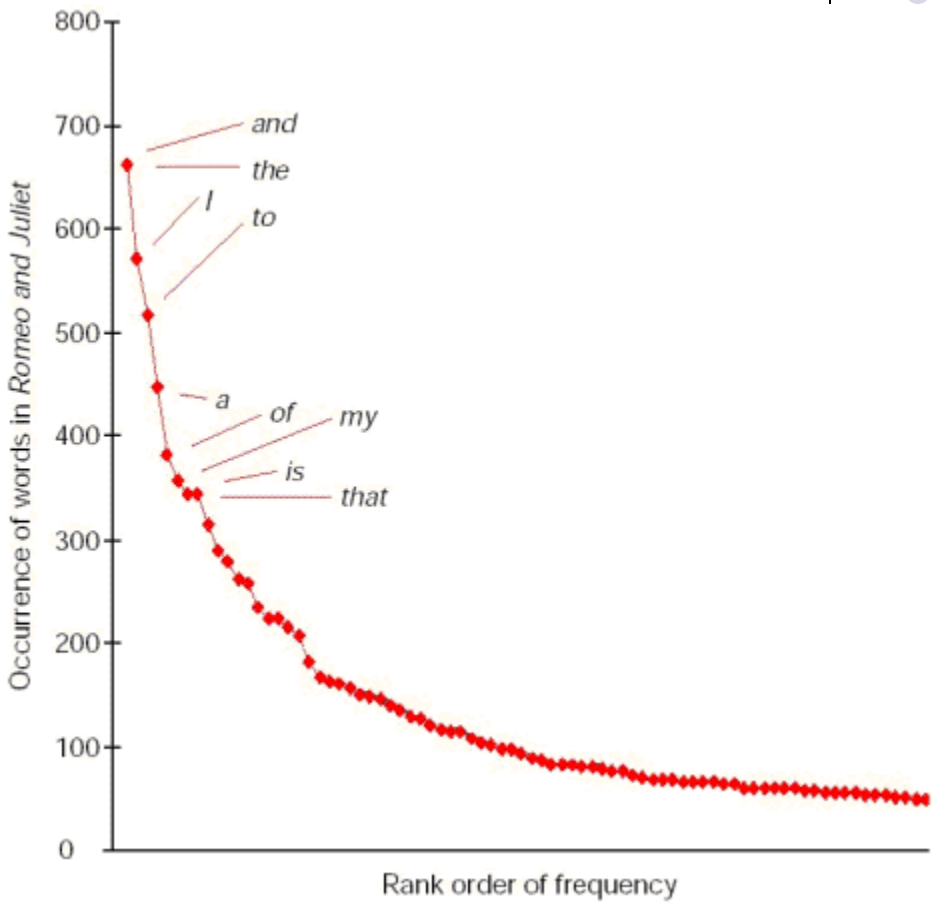
George K. Zipf
1902-1950

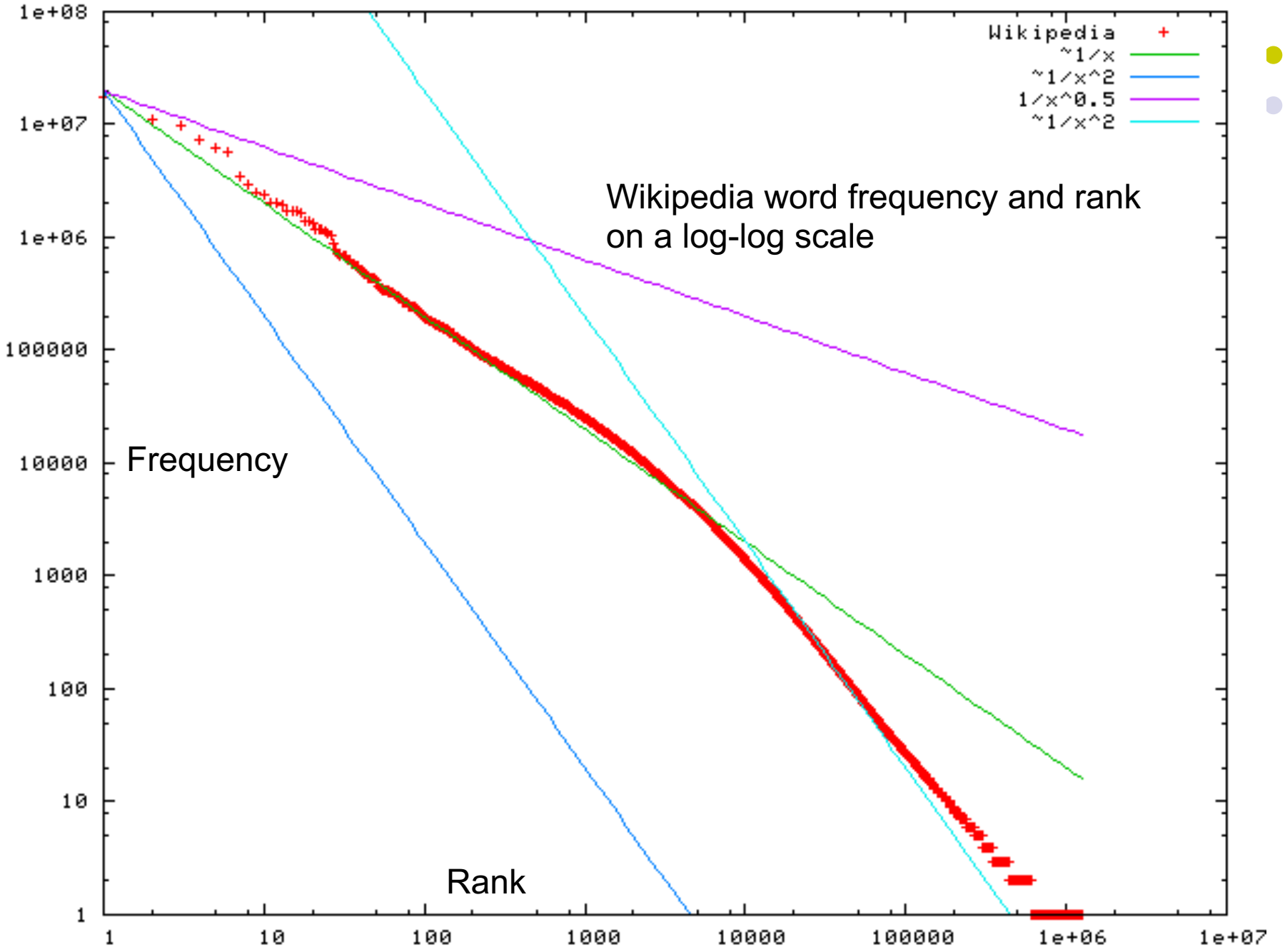
- Frequency of a word with rank r, $x_{(r)}$, is inversely proportional to its rank
  - $r^*x_{(r)} = c$ (constant)
  - E.g., what is the frequency of the 2nd ranked word in English relative to the first?
  - Does it really hold?

- A few words occur *very frequently*
- A medium number of word have, well, medium frequency!
- Many words occur *very infrequently*

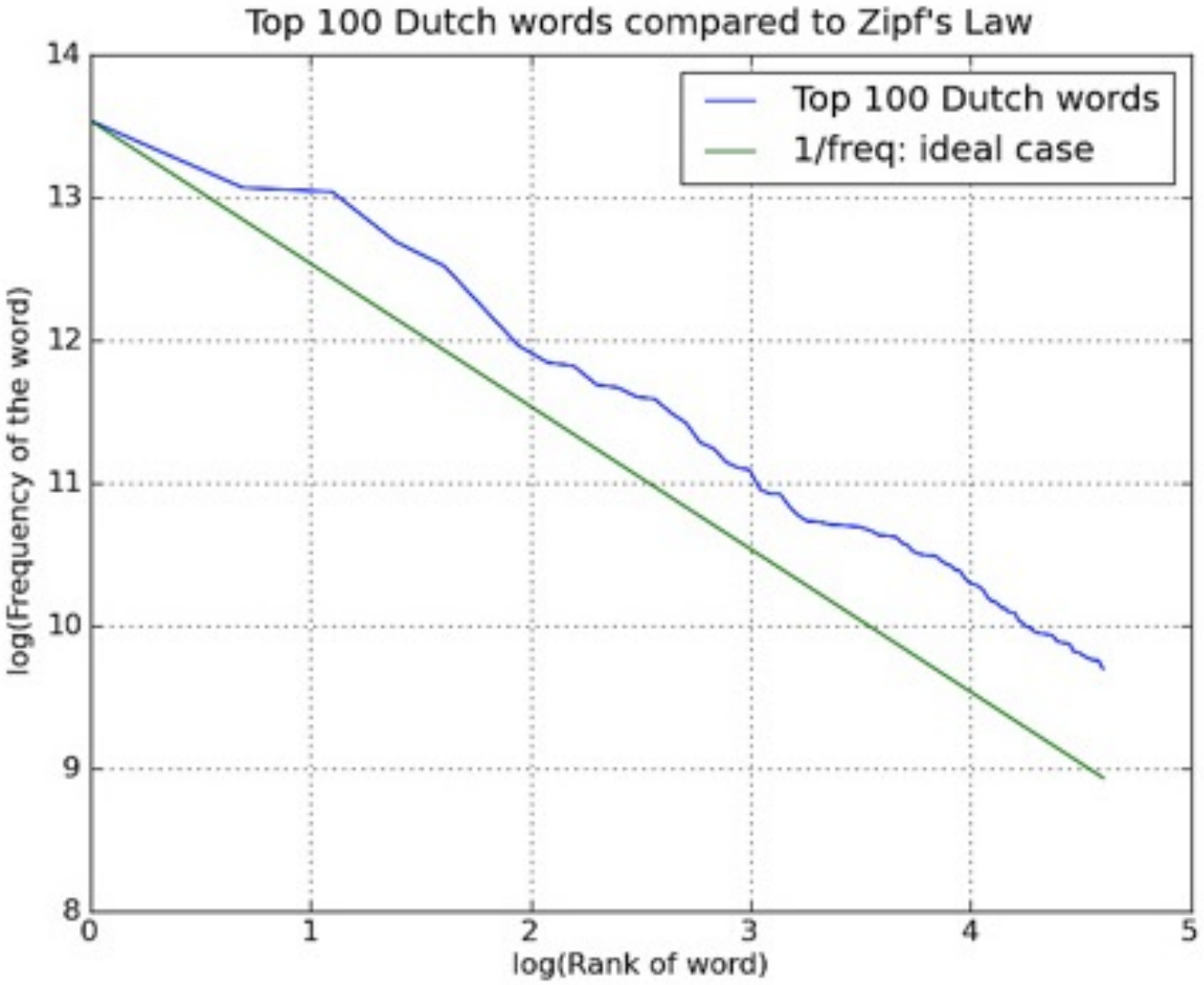
Word Distribution for Romeo & Juliet

The product of the frequency of words and their rank (*r*) is approximately constant (up to a certain rank)



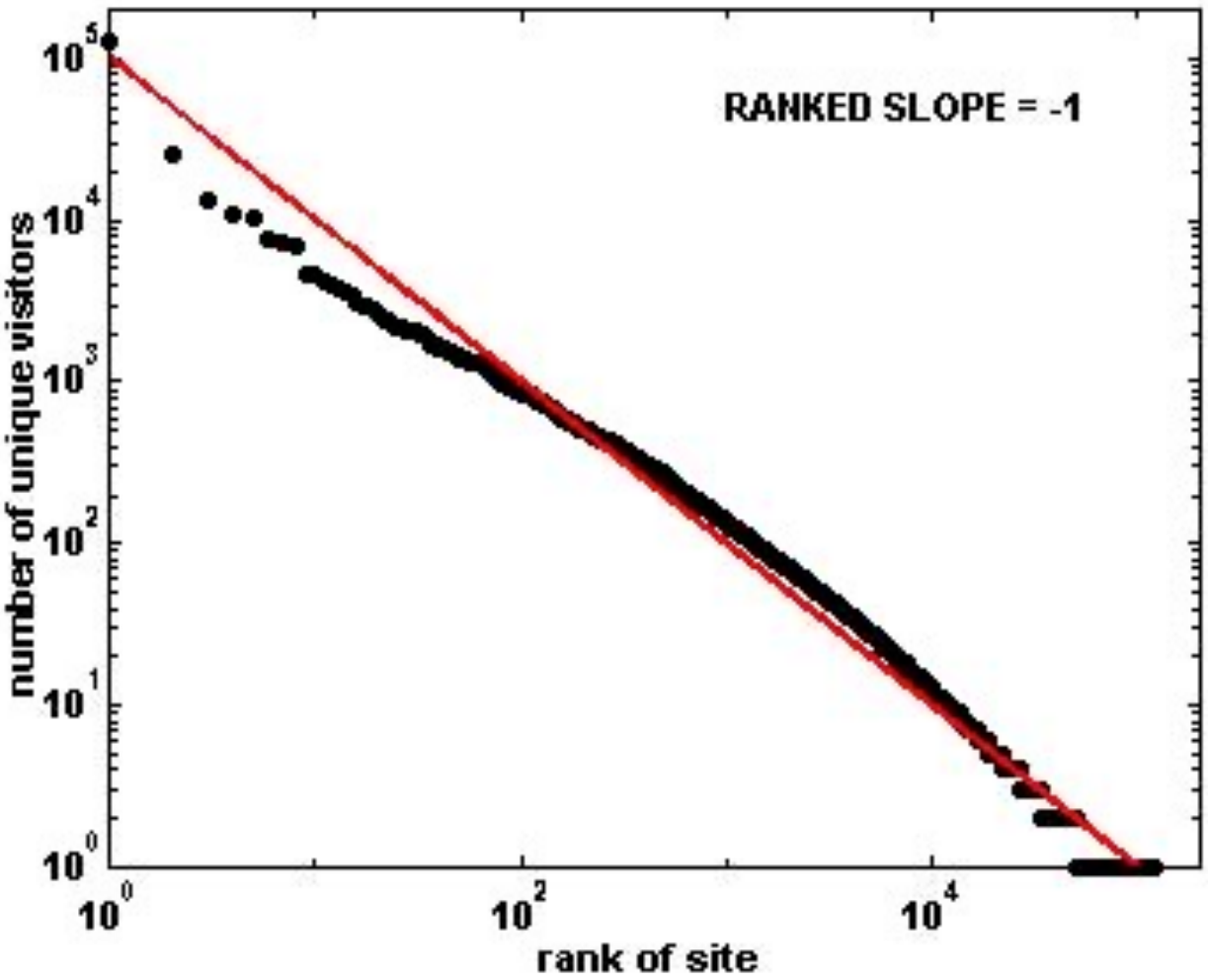


http://upload.wikimedia.org/wikipedia/commons/b/b9/Wikipedia-n-zipf.png

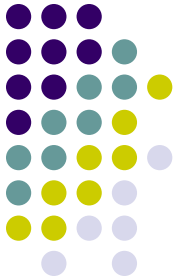


Zipf's law and top 100 Dutch words

Does Zipf's Law Fit This Distribution?



Hello Mr. Pareto!



X is a random variable with a Pareto distribution

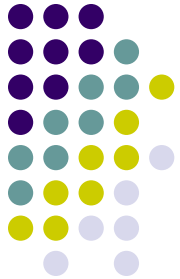
Survival function $P(X > x) = \frac{x_m^\alpha}{x^\alpha}$ if $x \geq x_m$, 1 otherwise

where $\alpha > 0$ and $x_m > 0$ is the minimum value of X

Cumulative distribution function $F_X(x) = 1 - \frac{x_m^\alpha}{x^\alpha}$ if $x \geq x_m$, 0 otherwise

Probability density function $f_X(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$ if $x \geq x_m$, 0 otherwise

Connection to Pareto's Law

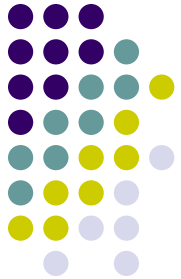


- Pareto's law: # events > x inversely related to a power of x
  - $P(X > x) \propto x^{-a}$
  - E.g., number of people with income > x
- Any correspondence with Zipf's law?

	Y-axis	X-axis
Zipf's law	# occurrences of words	Rank of word
Pareto's law	# people with income ≥ x	Income x

Source: http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html#ap2

Empirical Testing of Zipf's Law



$$\ln r = \beta_1 + \beta_2 \ln x_{(r)} + \varepsilon$$

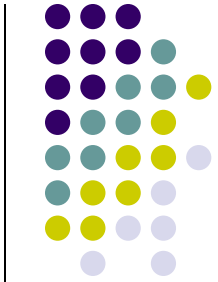
$r = 1, 2, \ldots, n$

Check if $\beta_2 = -1$

Limitations? How about testing for $\boldsymbol{\theta} = -1$ below?

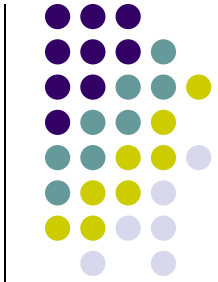
$$\ln r = \theta \ln \left(\frac{x_{(r)}}{n x_{(n)}} \right) + \varepsilon$$

From Unstructured to Structured Data



- Represent documents by attributes (e.g., presence or absence of terms)
- Can use existing data mining methods on attributes
- Vector representation
  - E.g., bag-of-words
- Add some statistical characteristics
  - E.g., term frequency, document frequency, length, etc.

Document Representation



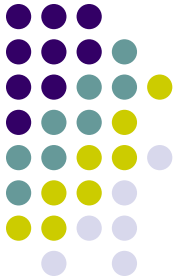
Binary representation: Presence of absence of terms

	Terms				
	Digital	Camera	Memory	Pixel	….
Doc 1	1	1	0	1	
Doc 2	1	1	1	0	
….					

But a term can occur multiple times in a document

	Terms				
	Digital	Camera	Memory	Pixel	….
Doc 1	1	2	0	1	
Doc 2	2	3	1	0	
….					

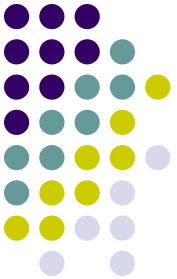
Using Relative Frequencies



Use # of occurrences of a term in a document / # terms in the document

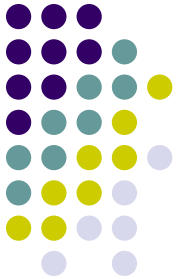
	Terms				
	Digital	Camera	Memory	Pixel	….
Doc 1	.01	.02	0	.01	
Doc 2	.002	.003	.001	0	
….					

Classification with Text



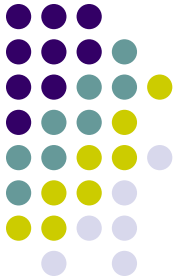
Text	Outcome (Class)
Senior position in corporate finance, 20+ years experience preferred	High salary
This restaurant was a waste of time & money. Everything sucks here!	Negative sentiment

But Not All Words Are Created Equal



- The main idea (for classification):
  - Words (or more generally, *terms*) that appear frequently in a document are …
  - Words that appear frequently in most documents are ….
  - Need a metric that shows the importance of a *term* to a *document* in a *corpus*.

Putting it Together

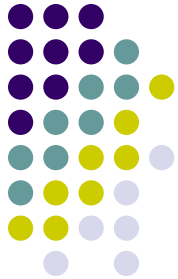


- Frequency of *j*th term in the *i*th document (call it $f_{i,j}$)
- Document frequency of the *j*th term = what % of documents have the term *(say,* F_j)
- *Term frequency-Inverse document frequency tf-idf =* $f_{i,j}$ * *log*(1/ F_j)

Term Frequencies	This	is	an	example	equation
Document 1	1	1	1	1	0
Document 2	1	1	1	0	1

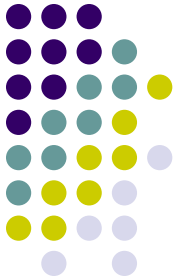
TF-IDF	This	is	an	example	equation
Document 1	1*log(2/2) = 0	1*log(2/2) = 0	0	1*log(2/1) = .3	0
Document 2	0	0	0	0	.3

Variations of TF



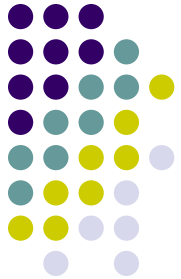
- Log normalized
  - $TF_{ij} = \log(1 + f_{ij})$
- Maximum normalization with damping
  - $Tf_{ij} = a + (1-a) * f_{ij} / max_i(f_{ij})$
  - where $max_i(f_{ij})$ is the highest term frequency in document i
  - $0 < a < 1$
- When would these variations be useful?

Some Issues in Text Classification



- The case of sentiment analysis
- Need to account for
  - Negation
    - E.g., *not* or *n't* followed by a verb can be treated as an additional feature
  - Parts of speech bigrams (esp. for sentiment analysis)
  - Trigrams, 4-grams, etc.
- The problem of adding *extra* features
  - An explosion of $|V|$
  - E.g., 25k reviews, 300k basic features, millions of features with bi- & tri-grams
  - Need to cut down
  - Thoughts?

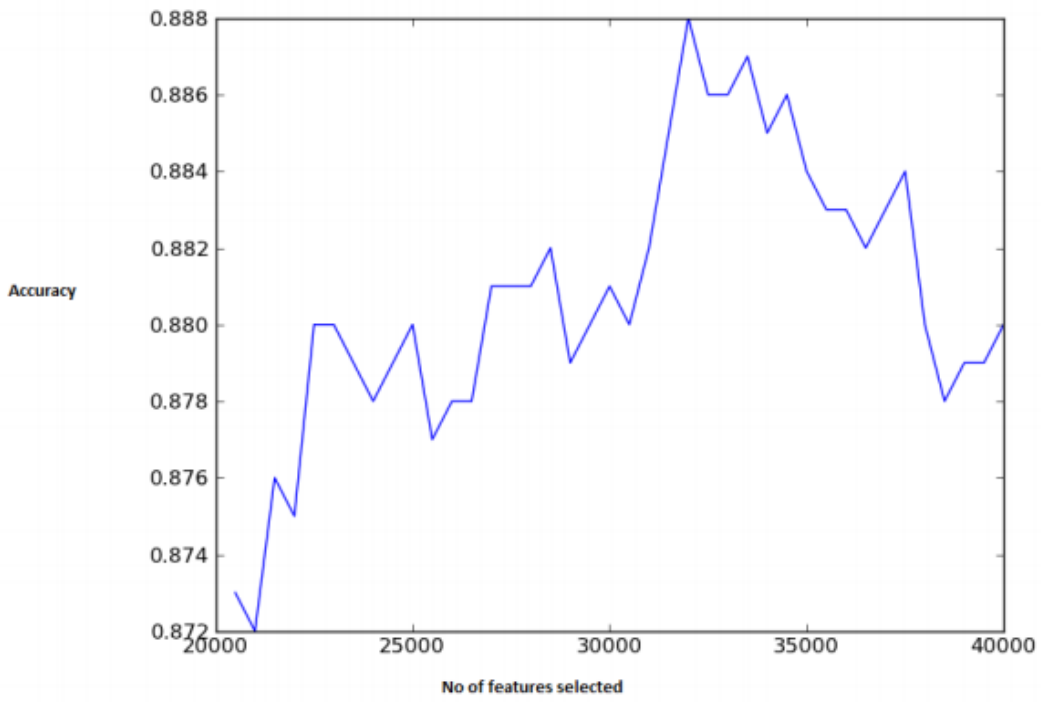
But the Vocabulary Becomes Too Large & Redundant



- Mutual information of two variables: The amount of information we have about one variable in the presence of the other.
- For discrete random variables *X* & *Y*:
- $MI(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} \left[p(x,y) * \log_2 \left(p(x,y)/(p(x)*p(y)) \right) \right]$
- No summation over X for words
- Binomial case
  - *x* = 0 or 1 (absence or presence of a feature)
  - *y* = document class (say, 0, 1)

Selecting Features

- For each feature, calculate its mutual information (with class)
- Plot classification accuracy as a function of the number of features selected, starting with features with highest mutual information



About 32k features with highest MI selected

Knowledge Discovery & Insights from User Generated Content

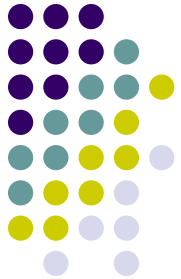
UNSTRUCTURED DATA ANALYTICS

MSITM, Session 2, 08/29/2022

Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor of Business
Distinguished Fellow, INFORMS Information Systems Society
University of Texas Distinguished Teaching Professor
McCombs School of Business, University of Texas at Austin
Email: aniteshb@gmail.com

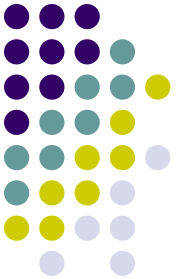
Some slides are based on “Mine Your Own Business Article”

The General Idea



- User generated content (UGC) can reveal new insights about brands, products & events
- The idea
  - Find out what people are talking about
  - Find key entities (e.g., brands, products, etc.) & attributes (e.g., service, price, etc.)
  - Find associations between words
  - Visualize, correlate with business outcomes (e.g., switching)

Co-occurrence of Terms



- Do terms or words appear together in messages by chance or due to real association?
  - E.g., is Volvo more likely to mentioned in conjunction with safety related words than Toyota?
  - Is Honda more likely to be associated with reliability than, say, Jaguar?

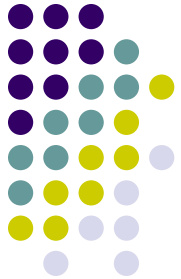
Let A and B represent two words or phrases

E. g. , $A = Volvo, B = safety$

$$Association\ or\ Lift(A,B) = \frac{p(A,B)}{p(A) * p(B)}$$

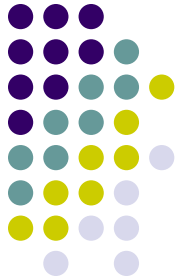
where $p(A,B)$ is the probability of both A & B appearing in a message
where $p(A)$ is the probability of A appearing in a message
where $p(B)$ is the probability of B appearing in a message

Interpretation of Lift



- *Lift* can be $= 1, > 1$ or < 1
- What does it mean when *Lift*$(A,B) = 1$?
- What about *Lift* $(A,B) > 1$?
- What is the significance of the word *Lift* in this context?
- What about *Lift*$(A,B) < 1$?
- From a practical standpoint we don't distinguish between *Lift*$(A,B) = 1$ or < 1

An Frequentist Interpretation of Lift



$$Let\ \#(A,B) = number\ of\ messages\ containing\ both\ A\ \&\ B$$

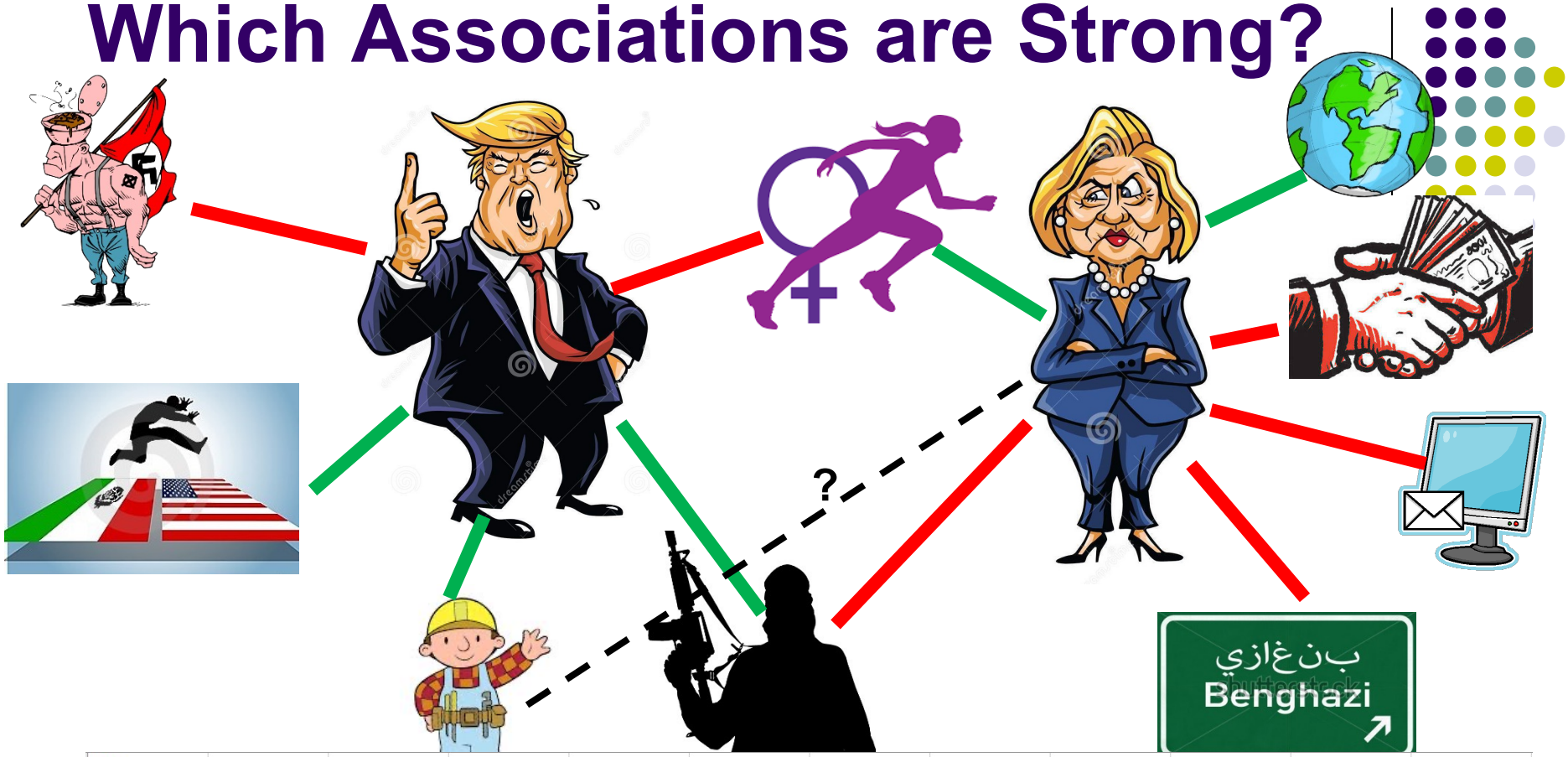
$$\#(A) = number\ of\ messages\ containing\ A$$

$$\#(B) = number\ of\ messages\ containing\ B$$

$$Then\ \ p(A,B) = \frac{\#(A,B)}{N}, \qquad p(A) = \frac{\#(A)}{N}, \qquad p(B) = \frac{\#(B)}{N}$$

$$So\ \ Lift(A,B) = \frac{\frac{\#(A,B)}{N}}{\frac{\#(A)}{N} * \frac{\#(B)}{N}} = \frac{N * \#(A,B)}{\#(A) * \#(B)}$$

Which Associations are Strong?



Lift											
	Trump	Clinton	Jobs	Email	Racism	Terrorism	Benghazi	Immigratio	Women	Global	Corruption
Trump		2.1	2.5	0.3	1.8	2.8	0.3	3.6	1.7	0.8	0.7
Clinton			0.7	2.7	0.3	1.3	2.7	0.7	2.1	2.4	1.6
Jobs				0.4	0.7	0.6	0.2	2.1	0.3	1.2	0.35
Email					0.75	0.45	1.8	0.1	0.25	0.57	1.1
Racism						0.8	0.1	0.89	0.7	0.4	0.2
Terrorism							1.6	1.2	0.4	1.05	0.6
Benghazi								0.2	0.35	0.7	0.65
Immigration									0.24	0.75	0.25
Women										0.4	0.2
Global											0.8
Corruption											

An Example from Edmunds.com

	Toyota Camry	Volvo S40
# mentions	34,559	1160
# co-mentions with safety related words	379	60
Lift of make/model & safety related words		

- Total # messages: 868,174 (assume one make/model appears only once in a message
- Safety related messages: 4534

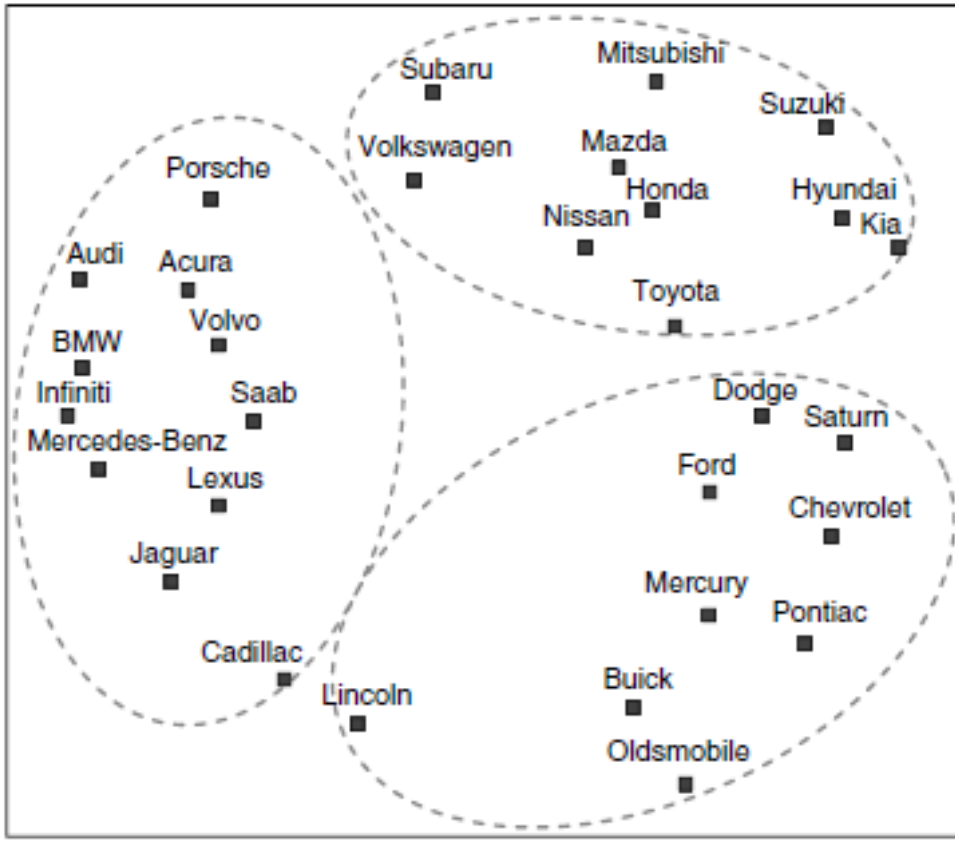
What are the values of *Lift*(*Volvo*, *safety*) and *Lift*(*Toyota*, *safety*)?

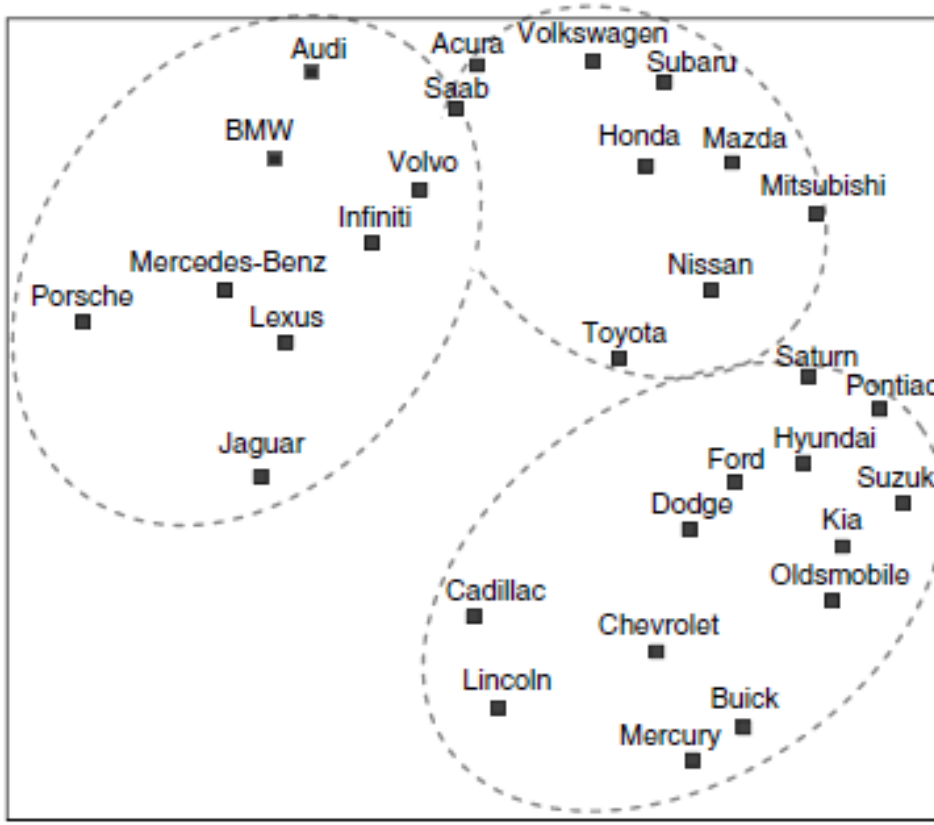
What Does This Ad Say?

Do car enthusiasts and customers feel the same way?

Should we Emphasize Both Performance & Luxury?

Can We Predict Switching Behavior?





Forum co-mentions
(100k mentions)

Actual brand switching
(3.5 million)

How to Draw an MDS Plot

Lift (Similarity)

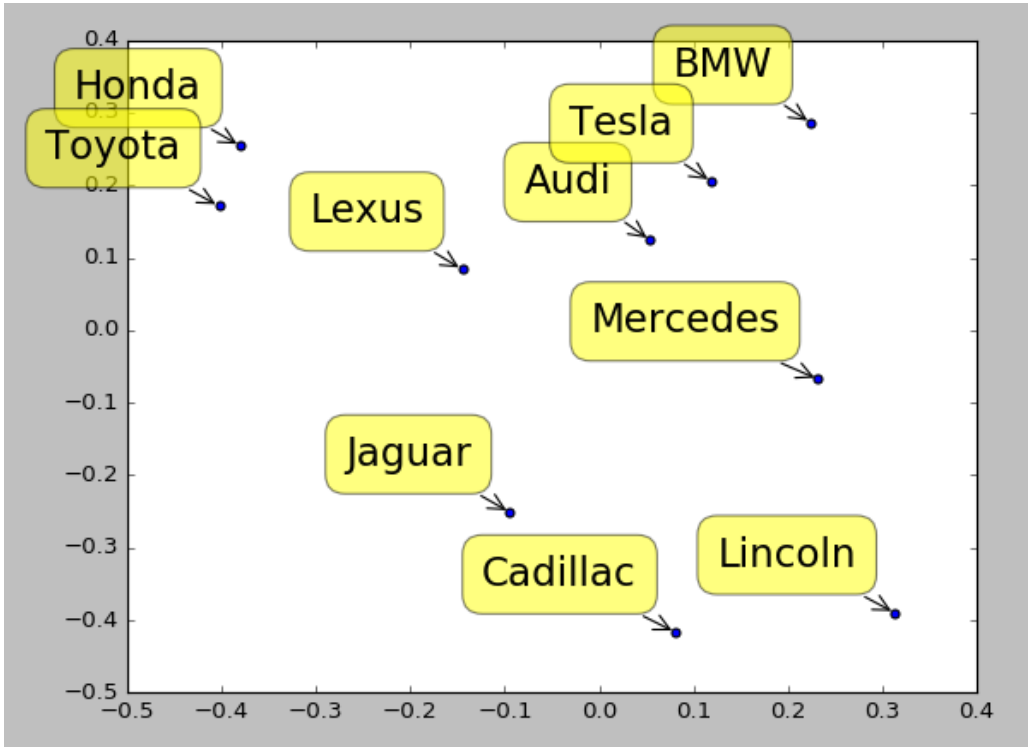
	Audi	BMW	Cadillac	Honda	Jaguar	Lexus	Lincoln	Mercedes	Tesla	Toyota
Audi		2.3	1.2	1.5	1.9	2.6	1.4	2.1	2.8	1.4
BMW			1.1	1.45	1.7	1.92	1.25	2.2	3.1	1.35
Cadillac				1.2	2.1	1.5	3.7	1.8	1.6	1.1
Honda					1.4	1.9	0.9	1.1	1.7	4.8
Jaguar						1.8	1.9	2	1.5	1.3
Lexus							1.2	2.4	2.1	2.9
Lincoln								2.2	1.3	0.8
Mercedes									2.3	1.4
Tesla										1.9
Toyota										

Should we plot the lift values directly?

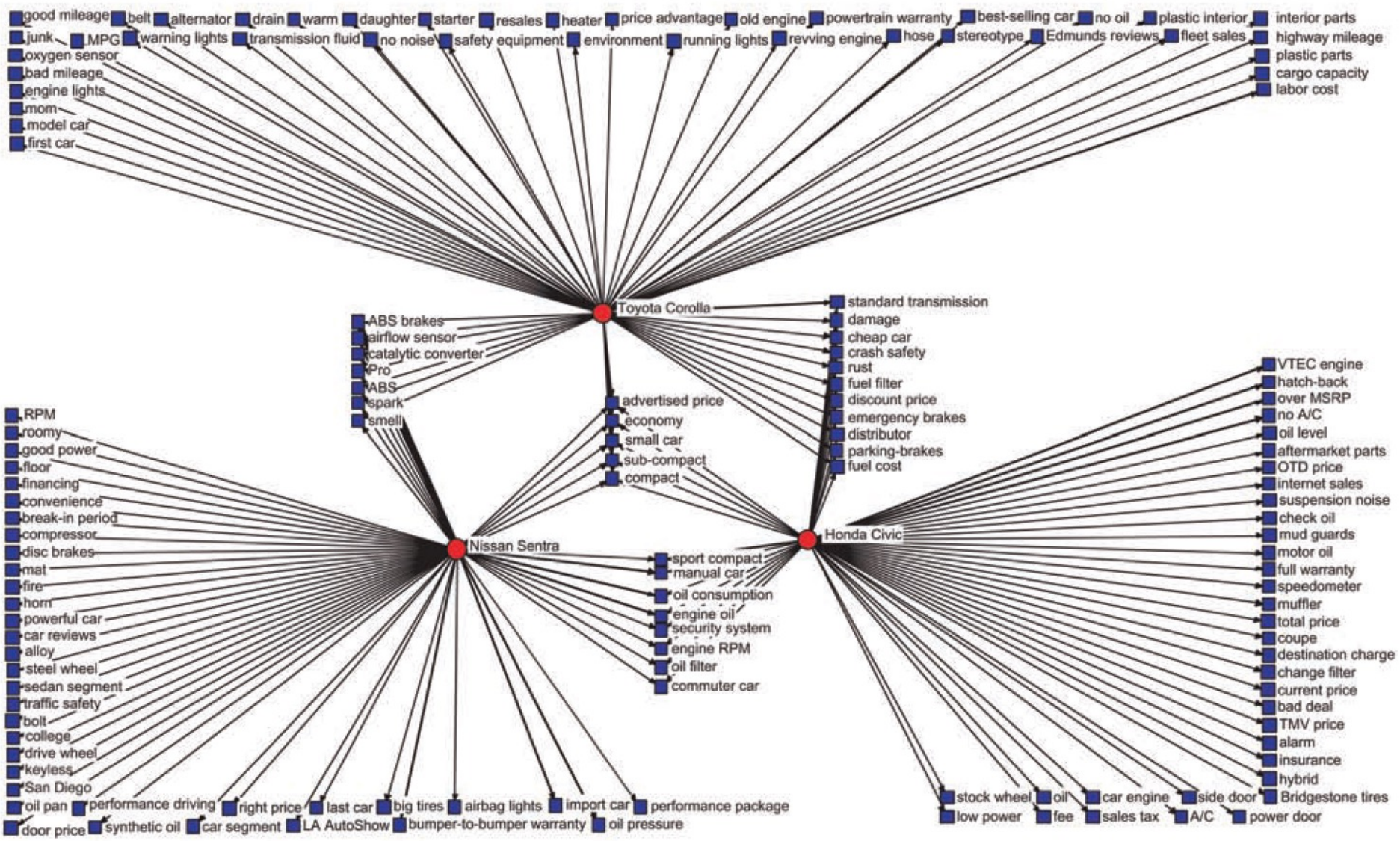
1/Lift (Dissimilarity)

	Audi	BMW	Cadillac	Honda	Jaguar	Lexus	Lincoln	Mercedes	Tesla	Toyota
Audi	0	0.434783	0.833333	0.666667	0.526316	0.384615	0.714286	0.47619	0.357143	0.714286
BMW		0	0.909091	0.689655	0.588235	0.520833	0.8	0.454545	0.322581	0.740741
Cadillac			0	0.833333	0.47619	0.666667	0.27027	0.555556	0.625	0.909091
Honda				0	0.714286	0.526316	1.111111	0.909091	0.588235	0.208333
Jaguar					0	0.555556	0.526316	0.5	0.666667	0.769231
Lexus						0	0.833333	0.416667	0.47619	0.344828
Lincoln							0	0.454545	0.769231	1.25
Mercedes								0	0.434783	0.714286
Tesla									0	0.526316
Toyota										0

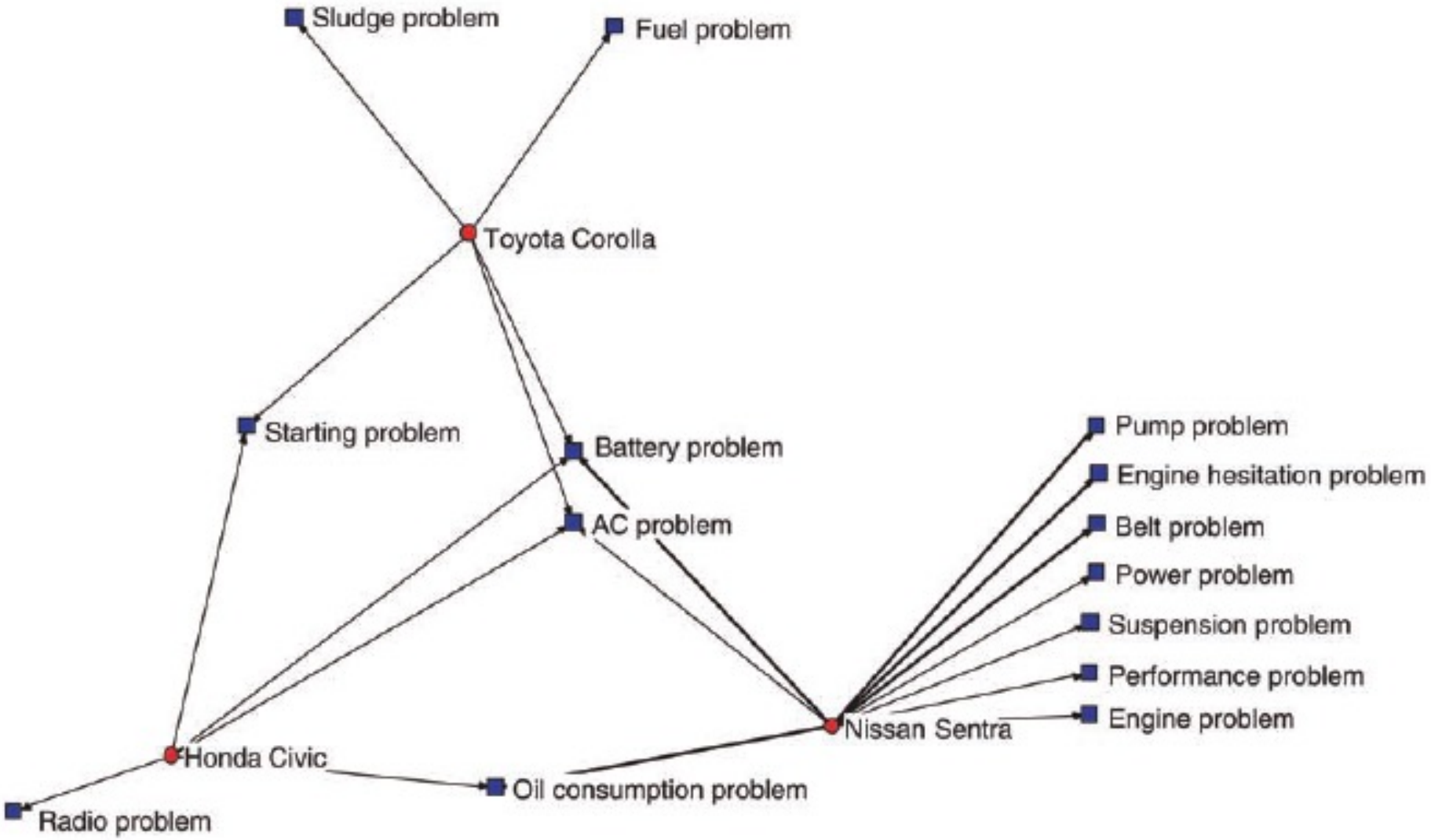
What do the MDS Axes Represent?



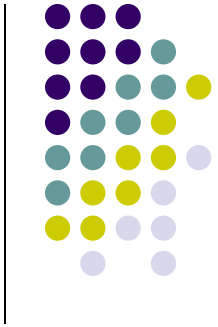
Product Attribute Associations



Negative Perceptions



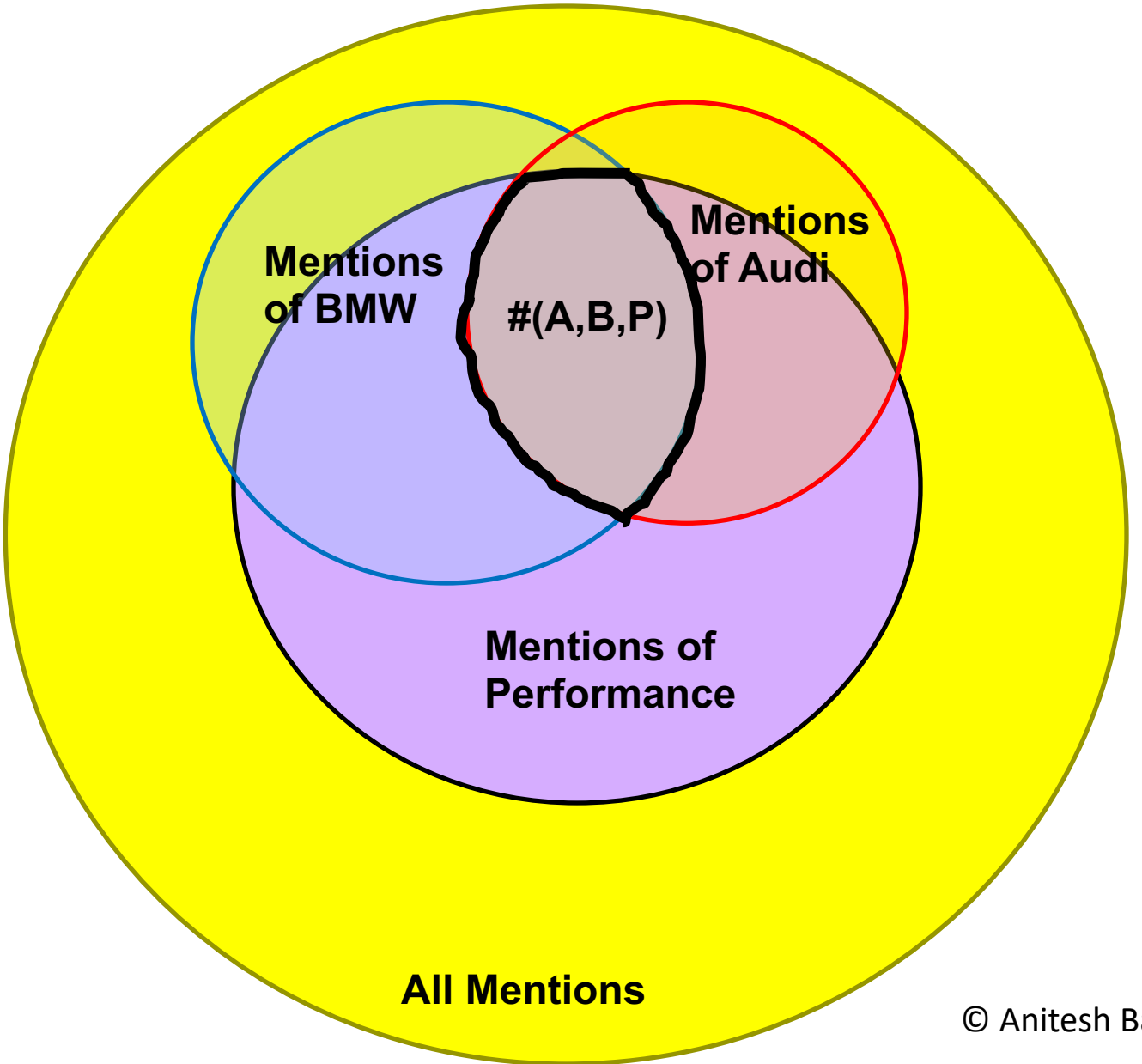
Crowds Vs. Experts



Forum mentions			WebMD	
Drug	ADR	Lift	Frequency	Severity
Actos	Fluid retention	6.51	Infrequent	Severe
Actos	Liver problems	4.89	Rare	Severe
Actos	Edema	4.54	Rare	Severe
Actos	Swelling	4.45	Infrequent	Severe
Actos	Weight gain	3.12	Rare	Severe
Amaryl	Low blood sugar	8.23	Infrequent	Severe
Amaryl	Weight gain	3.81	Doesn't exist	
Avandia	Heart problems	6.77	Rare	Severe
Avandia	Edema	6.42	Rare	Severe
Avandia	Swelling	4.25	Infrequent	Severe
Avandia	Fluid retention	3.31	Infrequent	Severe
Avandia	Weight gain	2.24	Rare	Severe
Byetta	Bad taste	2.87	Rare	Less severe
Byetta	Hair loss	2.86	Rare	Less severe
Byetta	Jitteriness	2.55	Infrequent	Less severe
Byetta	Nausea	2.46	Common	Less severe
Byetta	Loss of appetite	2.42	Infrequent	Less severe
Byetta	Cold symptoms	2.35	Doesn't exist	
Byetta	Constipation	2.22	Rare	Less severe
Byetta	Bloated feeling	1.83	Rare	Less severe
Byetta	Rash	1.72	Rare	Severe
Glucotrol	Low blood sugar	4.42	Common	Severe
Glyburide	Increased hunger	5.45	Common	Less severe
Glyburide	Weight gain	2.59	Common	Severe
Humalog	Allergic reaction	8.92	Common	Severe
Humalog	Rapid heartbeat	6.84	Rare	Severe
Humalog	Kidney problems	5.58	Doesn't exist	
Januvia	Respiratory problems	10.59	Infrequent	Severe
Januvia	Jitteriness	9.27	Rare	Severe
Januvia	Irritability	6.18	Rare	Severe
Januvia	Sinus problems	5.29	Infrequent	Severe
Januvia	Cold symptoms	3.13	Infrequent	Less severe
Lantus	Mood problems	9.78	Doesn't exist	
Lantus	Irritability	5.25	Rare	Severe
Lantus	Lower blood sugar	2.90	Common	Severe
Levemir	Anxiety problems	9.34	Doesn't exist	
Levemir	Sleep problems	8.49	Doesn't exist	
Levemir	Allergic reaction	6.14	Rare	Severe
Levemir	Rash	3.70	Infrequent	Severe
Metformin	Lactic acid	3.76	Rare	Severe
Metformin	Taste problems	3.76	Common	Less severe
Metformin	Muscle pain	2.88	Infrequent	Less severe
Metformin	Stomach cramps	2.76	Common	Less severe
Metformin	Diarrhea	2.49	Common	Less severe
Metformin	Digestive disorders	2.49	Common	Less severe
Metformin	Leg pain	2.07	Infrequent	Less severe
Symlin	Low blood sugar	5.78	Infrequent	Severe
Symlin	Bloated feeling	3.62	Doesn't exist	
Symlin	Nausea	1.80	Common	Less severe

Beyond Pairwise Association

Do people compare **Audi** & **BMW** when they discuss **performance**?



Conditional Lift

- $Lift(x,y|z) = P(x,y|z)/[P(x|z)*P(y|z)]$

$= [\#(x,y,z)/\#(z)] / [\{\#(x,z)/\#(z)\}*\{\#(y,z)/\#(z)\}]$

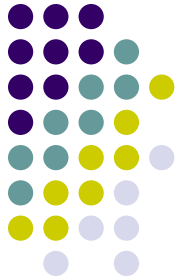
$= [\#(x,y,z)*\#(z)] / [\#(x,z)*\#(y,z)]$

- E.g., L(Audi, BMW|performance) =

[#(Audi, BMW, performance)*#(performance)] /
[#(Audi,performance)*#(BMW,performance)]

- Perf: 500; Audi & perf: 250; BMW & perf: 300; Audi, BMW & Perf: 200
- $L(A,B|P) = (200*500)/(250*300) = 1.33$

How is Lift Different from Confidence?



- Commonly found in commercial software packages
- n% of people who mentioned word x (e.g., BMW) also mentioned y (e.g., Lexus)
- $Confidence(Lexus \mid BMW) = \#(Lexus, BMW)/\#(BMW) =$ (say) $125 / 300 = .417$
- Confidence is not symmetric
- Also if most people talk about Lexus anyway, $Confidence(Lexus \mid BMW)$ will not be a useful metric.

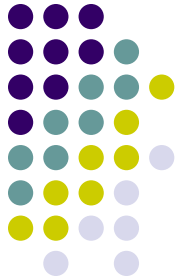
Unstructured Data Analytics

Network Value of a Customer
Homophily

MSITM, Fall 2022, Nov 14

Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor of Business
Distinguished Fellow, INFORMS Information Systems Society
Stevens Piper Foundation Professor
University of Texas Distinguished Teaching Professor
Associate Director, Center for Research in e-Commerce
McCombs School of Business, University of Texas at Austin
Email: aniteshb@gmail.com

The Network Value of a Customer



- The distinction between customer lifetime value (CLV) and customer network lifetime value (CNLV)
- Not a new concept
  - Celebrity endorsements
  - E.g., maybe very low CLV but possibly very high CIV
- So what has changed and why?

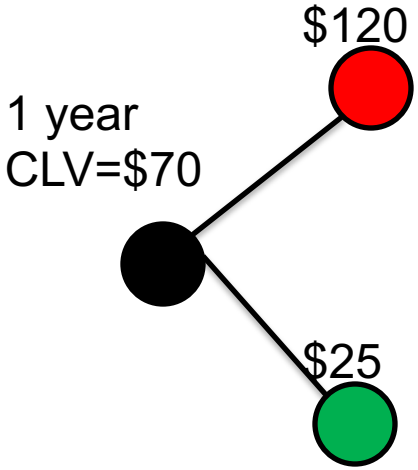
Source: A. Azhar, www.peerindex.com

High CLV and CIV May Not Go Together

	Low CIV	High CIV
High CLV	"Affluent"	"Champion"
Low CLV	"Irrelevant"	"Advocate"

Adapted from Kumar, Petersen & Leone

Customer Influence Value Example



Would not buy without influence

Would buy anyway with promotion

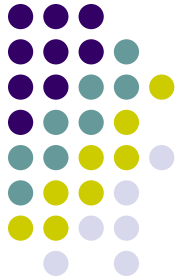
- Average contribution margin = $100
- Average acquisition cost = $25
- Retention cost per year = $5
- A customer influences one "red" and one "green" user

- Value of influencing type "red" = 100 + 25 – 5 = $120
- Value of influencing type "green" = $25 (saving the acquisition cost)

1-year *CIV* (without discounting) = $145

How can we tell the type?

Identifying the “Red” & “Green” Types



- Use data on
  - Demographics & transactions
  - Acquisition success/failure in the past
- To develop red & green classification model
- New customer acquired through influence/word-of-mouth
- Classify as type red or green
- Calculate the value of your influencers

Where is the Return on Influence (ROI)?

- Can social or digital Word-of-Mouth (WOM) ROA be higher than that of paid search?
- What kind of incentives are required for spreading the word in social networks?
- How can we measure social ROA?

Image Source: Mark Jeffery, Kellogg School

Direct Vs. WOM Traffic

Rank by Direct Traffic	Site	Direct Referrals
1	CAMPAIGN SITE	14,467
2	ad.adlegend.com (AD SERVER)	12,850
3	g.doubleclick.net (AD SERVER)	8,611
4	www.jeuxvideo.com	7,844
5	www.youtube.com	5,412
6	FAN SITE	4,455
7	forums.gametrailers.com	3,678
8	es.wikipedia.org	3,630
9	FAN SITE	3,494
10	[redacted]	2,251
11	www.meristation.com	2,247
12	answers.yahoo.com	2,064
13	mail.live.com	1,985
14	www.giga.de	1,906
15	COMPANY SITE	1,650
16	www2.hshare.net	1,531
17	www.spaziogames.it	1,481
18	www.akiba-online.com	1,477
19	www.joystiq.com	1,097
20	www.neogaf.com	1,045
21	FAN SITE	1,026
22	www.xbox360achievements.org	725
23	CAMPAIGN SITE	291
24	es.youtube.com	72
25	www.jeuxactu.com	61

The common view of traffic sources

WOM (Social) Referrals

Rank Direct + WOM traffic	Rank Direct traffic	SITE	Direct traffic	WOM traffic	LIFT: WOM/Direct
1	1	CAMPAIGN SITE	14,467	2,826	20%
2	9	FAN SITE	3,494	13,780	394%
3	21	FAN SITE	1,026	15,302	1491%
4	7	forums.gametrailers.com	3,678	11,958	325%
5	22	www.xbox360achievements.org	725	14,656	2022%
6	2	ad.adlegend.com (AD SERVER)	12,850	247	2%
7	15	FAN SITE	1,650	11,060	670%
8	4	www.jeuxvideo.com	7,844	2,634	34%
9	3	g.doubleclick.net (AD SERVER)	8,611	86	1%
10	20	www.neogaf.com	1,045	7,112	681%
11	5	www.youtube.com	5,412	1,287	24%
12	6	FAN SITE	4,455	731	16%
13	8	es.wikipedia.org	3,630	1,005	28%
14	23	FAN SITE	291	3,611	1241%
15	24	es.youtube.com	72	2,500	3472%
16	11	www.meristation.com	2,247	131	6%
17	10	[redacted]	2,251	13	1%
18	25	www.jeuxactu.com	61	2,171	3559%
19	13	mail.live.com	1,985	219	11%
20	12	answers.yahoo.com	2,064	11	1%
21	19	www.joystiq.com	1,097	967	88%
22	14	www.giga.de	1,906	16	1%
23	16	www2.hshare.net	1,531	48	3%
24	17	www.spaziogames.it	1,481	63	4%
25	18	www.akiba-online.com	1,477	2	0%

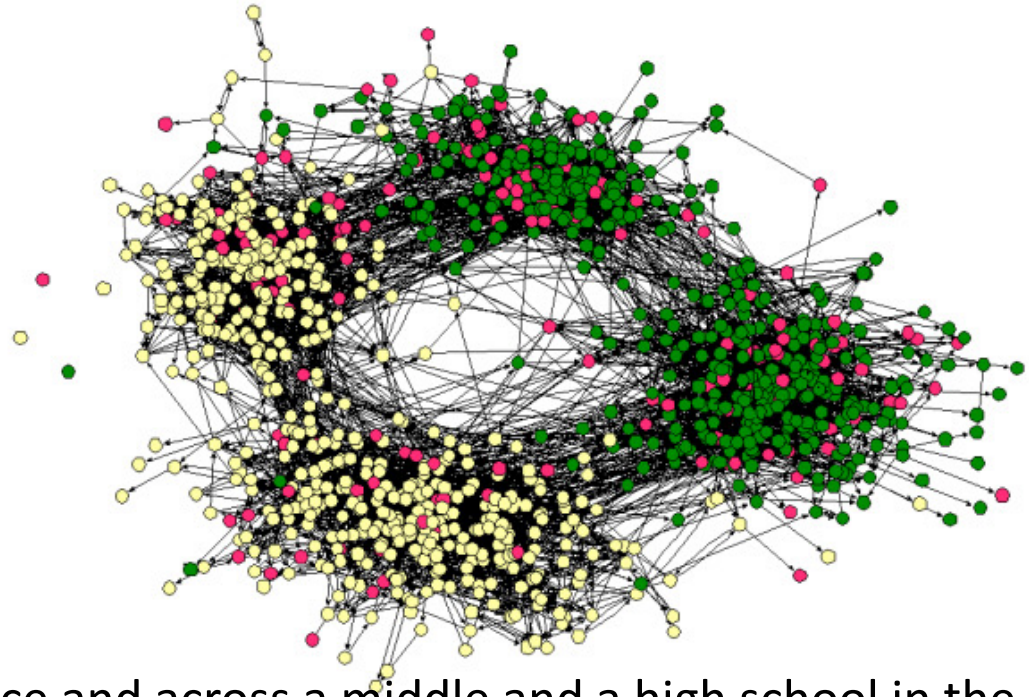
Meteor provides data required to identify the most influential sites.It's not always who you might think…

“If You Got the Money Honey I Got the Time”

		Marketing Tactic				
Traffic	Metric	Paid search	Organic search	Display ads	Email	Total
Direct	Total cost	$9,342	$3,000	$10,428	$2,000	$24,770
	Customer value	$50	$50	$50	$50	$50
	Direct visits	12,039	23,637	9,472	1,294	34,532
	Conversions	473	1,064	313	83	1933
	Conversion rate	0.039	0.045	0.033	0.064	0.056
	Direct profit	$14,308	$50,200	$5,222	$2,150	$71,880
WOM	WOM Visits	1960	7879	937	609	11385
	Conversions	136	697	93	110	1036
	Conversion rate	0.069	0.088	0.099	0.181	0.091
	WOM profit	$6,800	$34,850	$4,650	$5,500	$51,800
	Profit lift = WOM $ / Direct $	0.4752586	0.694223108	0.89046	2.5581	0.72065

Homophily (Similarity)

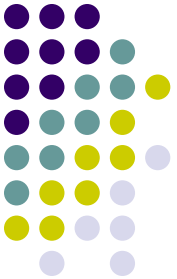
- "Birds of a feather flock together"
  - Your friends/contacts vs. a random sample of people
  - Social networks tend to connect people who are similar to each other



Friendships by race and across a middle and a high school in the same school district

Source: Easley & Kleinberg

Distinguishing Between Social Influence and Homophily

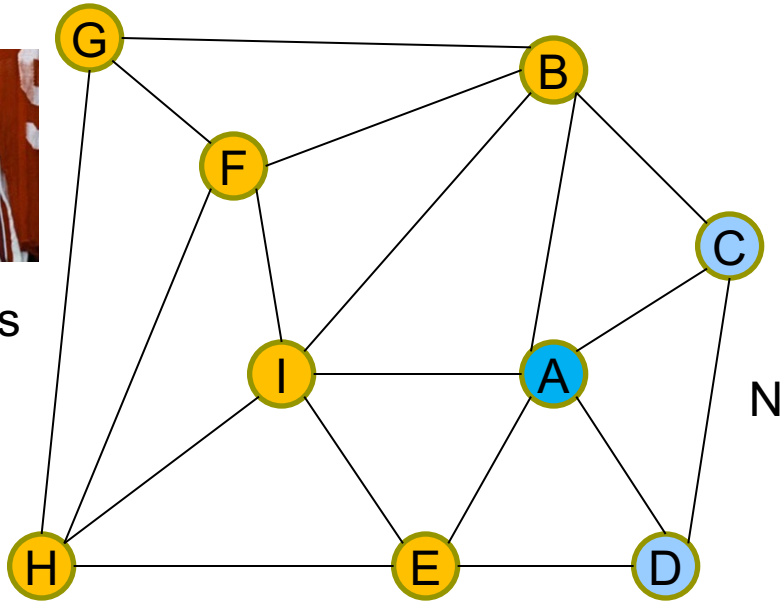


- Can opinions, attitudes & purchases be attributed to social influence?
- Or is it due to homophily?
- E.g., did I buy something because
  - you influenced me?
  - we are just similar?
- What difference would it make to a company's strategy?

Detecting Homophily for Static Attributes

- Have to know what attribute(s) may be relevant
- E.g., gender, interest, educational background, etc.

Jocks



Nerds

- Does this network exhibit homophily?
- What measure can we use?

Source: Easley & Kleinberg

A Little Theoretical Detour

A network with a set of nodes (V) & randomly assigned edges (E^r):

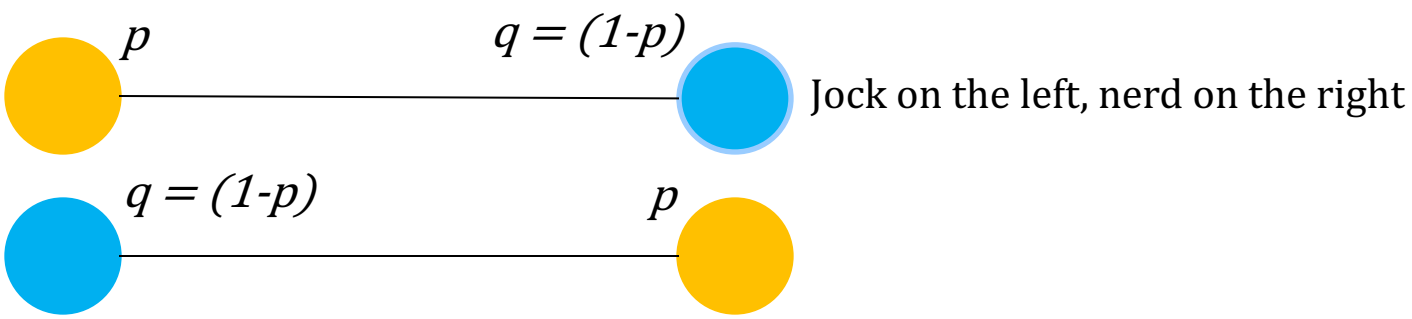
$$R = (V, E^r)$$

Each node is assigned an attribute: say, type = jock with probability p, and type = nerd with probability $q = 1\text{-}p$

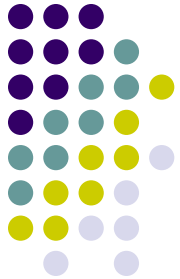
Consider any edge $(i,j) \in E^r$ of this random network R.

Let the random variable $X_{ij} = 1$ if it is a "cross-edge", and $X_{ij} = 0$ otherwise. Then X_{ij} is a Bernoulli random variable such that

$P(X_{ij} = 1) = 2pq$



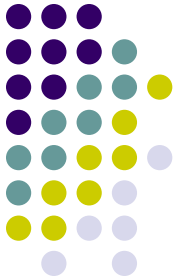
Dynamic Attributes: How Can We Distinguish Between Homophily & Influence?



- Need multiple snapshots in time
- Homophily: Due to **similar** attributes in time *t*, some people may choose to become friends in *t*+1
  - E.g., high achievers in a class may form links
- But some people may become friends in *t*+1 even though their attributes were different in *t*
- Check which effect is stronger

Source: LaFond & Neville

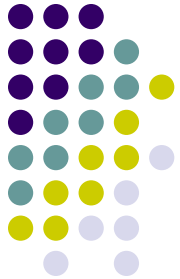
Test for Homophily



- Homophily exists if
- *p*(Becoming friends in *t*+1 where attributes were same in *t*) > *p*(Becoming friends in *t*+1 where attributes were different in *t*)
- *p*(Dissolving friendships in *t*+1 where attributes were same in t) < *p*(Dissolving friendships in *t*+1 where attributes were different in *t*)

Source: LaFond & Neville

Detecting Social Influence



- Some **friends** at *t* with different attributes may become similar in *t*+1 (due to social influence)
  - E.g., some buy a product their friends have
  - Some change their beliefs & attitudes
- But people who are not friends and have different attributes at *t* can also become similar at *t*+1 due to "other" factors
- Which effect is stronger?
  - I.e., is *p*(Attributes becoming same in *t*+1 where the individuals were friends in *t*) > *p*(Attributes becoming same in *t*+1 where the individuals were not friends in *t*)?
  - Is *p*(Attributes becoming different in *t*+1 where the individuals were friends in *t*) < *p*(Attributes becoming different in *t*+1 where the individuals were not friends in *t*)?

Source: LaFond & Neville

Distinguishing Between Homophily & Social Influence: The Case of Dynamic Attributes

Time: *t* *t*+1

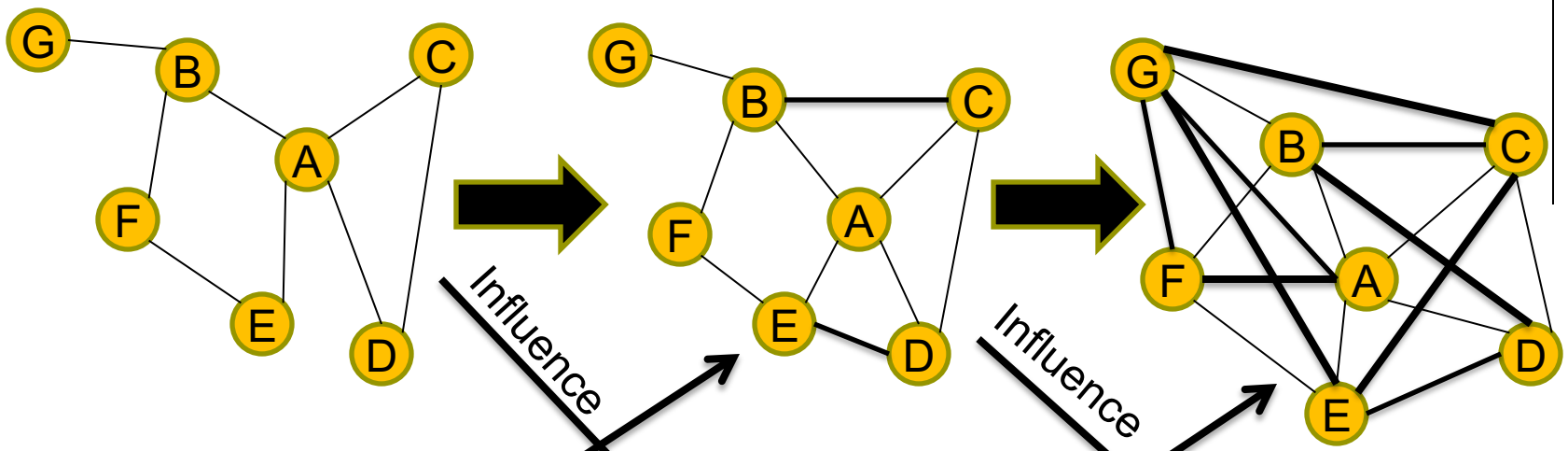
	Subscription
A	Yes
B	No
C	No
D	Yes
E	No
F	Yes
G	Yes

Attribute (e.g., subscription to a music service at time *t*

A	Yes
B	Yes
C	Yes
D	Yes
E	No
F	Yes
G	Yes

Attribute at time t+1

Evidence of Homophily Versus Social Influence



Time: *t*

t+1

t+2

Influence

Homophily

Influence

Homophily

A	Yes
B	No
C	No
D	Yes
E	No
F	Yes
G	Yes

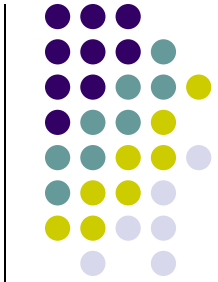
Attribute at time *t*

A	Yes
B	Yes
C	Yes
D	Yes
E	No
F	Yes
G	Yes

Attribute at time *t*+1

Attribute at time *t*+2

Testing Significance Levels



X^i is an atttribute of node v_i

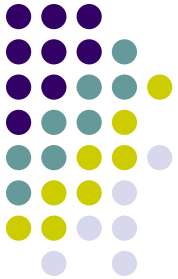
P_R is a set of related nodes (friends) in a network

	$X^i = X^j = x$	$\neg\,(X^i = X^j = x)$
$(v_i, v_j) \in P_R$	a	b
$(v_i, v_j) \notin P_R$	c	d

$$Relational\ autocorrelation\ C(X,G) = \chi^2 = \frac{N.(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Source: LaFond & Neville

Chi-Square (for Contingency Tables)



	Improved Outcome	Didn't improve	Total
Treatment	36	14	50
No Treatment	30	25	55
Total	66	39	105

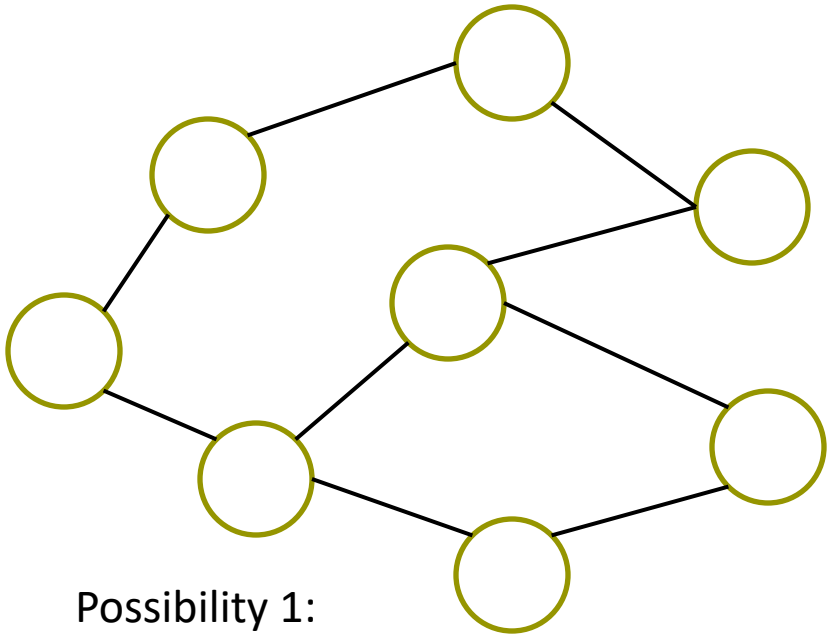
$$\chi^2 = ?$$

$$\chi^2 = \frac{105*(36*25-30*14)^2}{50*55*66*39} = 3.42$$

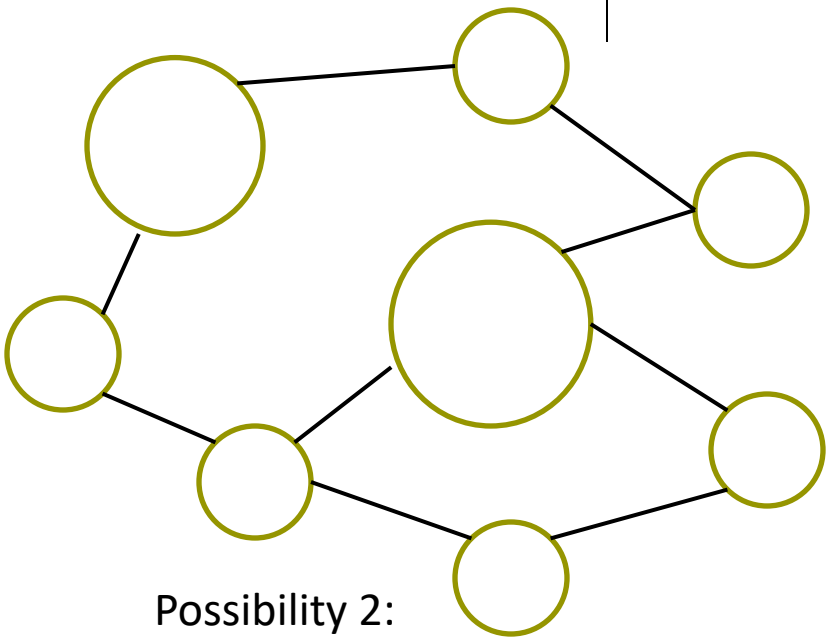
Significance Tests

- Homophily: $C(X_t, G_{t+1}) > C(X_t, G_t)$
- Social influence: $C(X_{t+1}, G_t) > C(X_t, G_t)$

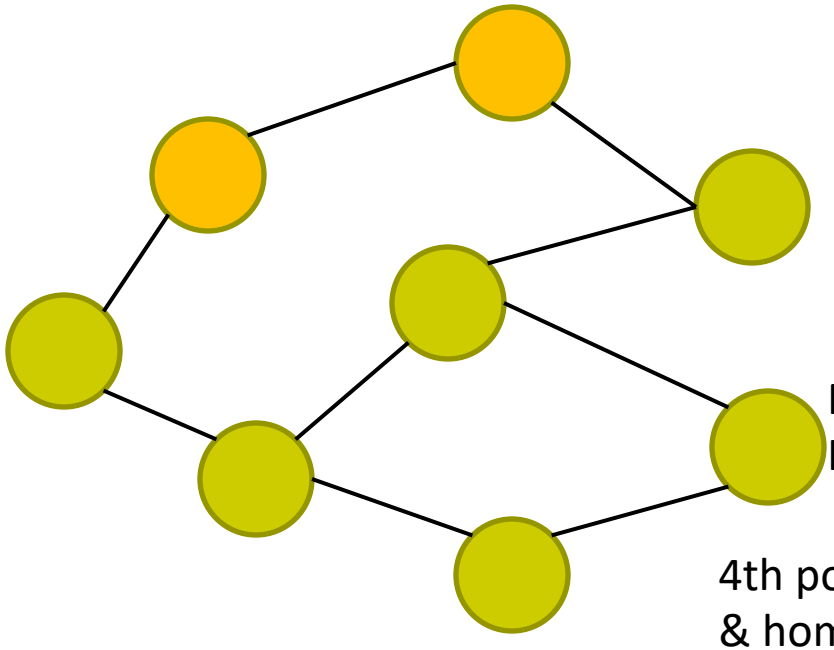
Not all Networks are Created Equal



Possibility 1:
No social influence, no homophily



Possibility 2:
Social influence but no homophily



Possibility 3:
Homophily but no social influence

4th possibility (not shown): Both influence & homophily

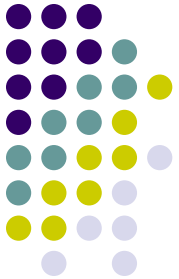
UNSTRUCTURED DATA ANALYTICS

MSITM
Session 1, 08/22/2022

Dr. Anitesh Barua
David Bruton Jr. Centennial Chair Professor of Business
Distinguished Fellow, INFORMS Information Systems Society
University of Texas Distinguished Teaching Professor
McCombs School of Business, University of Texas at Austin
Email: aniteshb@gmail.com

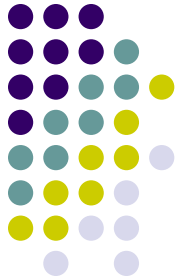
Course TA: Sruthi Pilla (sruthipilla@utexas.edu)

Learning Objectives



- Course details
- Why bother about unstructured data
- Contemporary business applications & use cases

A Serious Analytics Course, But What is the Connection to Business?



- Which brands do people compare with entry level Mercedes cars?
- How can we predict which brands and models customers will switch from to, say, a Tesla?
- How can we recommend products that meet specific customer requirements?
- How can we predict the sales of a new product that is yet to be released?
- How does Netflix figure out what will be a hit series?
- How does Airbnb predict which images and descriptions get a property rented faster?
- How can we discover new side effects of a prescription drug?

- Yes, it connects to business heavily!

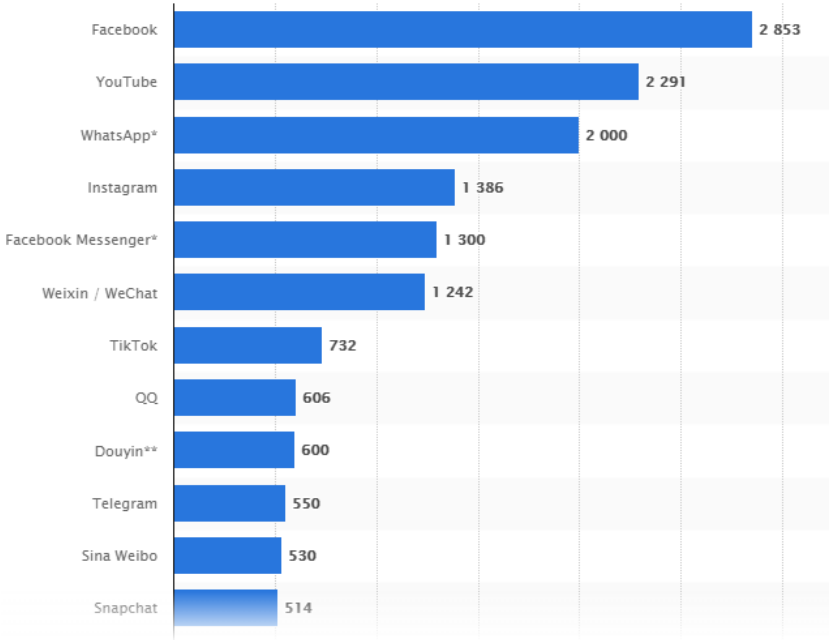
User Generated Content Analytics

- User Generated Content (UGC) = unstructured data (text, images, audio, video), structured data is numeric

- Two sources of UGC
  - External: Social media, public documents (e.g., shareholder reports)
  - Internal: HR evaluations, contracts, vendor reports, customer conversations, maintenance reports, medical reports, etc.

Why Bother About Unstructured Data?

- > 80% of all data are unstructured
- Web pages, online conversations/posts, digital libraries, etc.
- Twitter: 600 million daily tweets
- FB: 5+ billion pieces of content shared daily
- Conversations in online forums

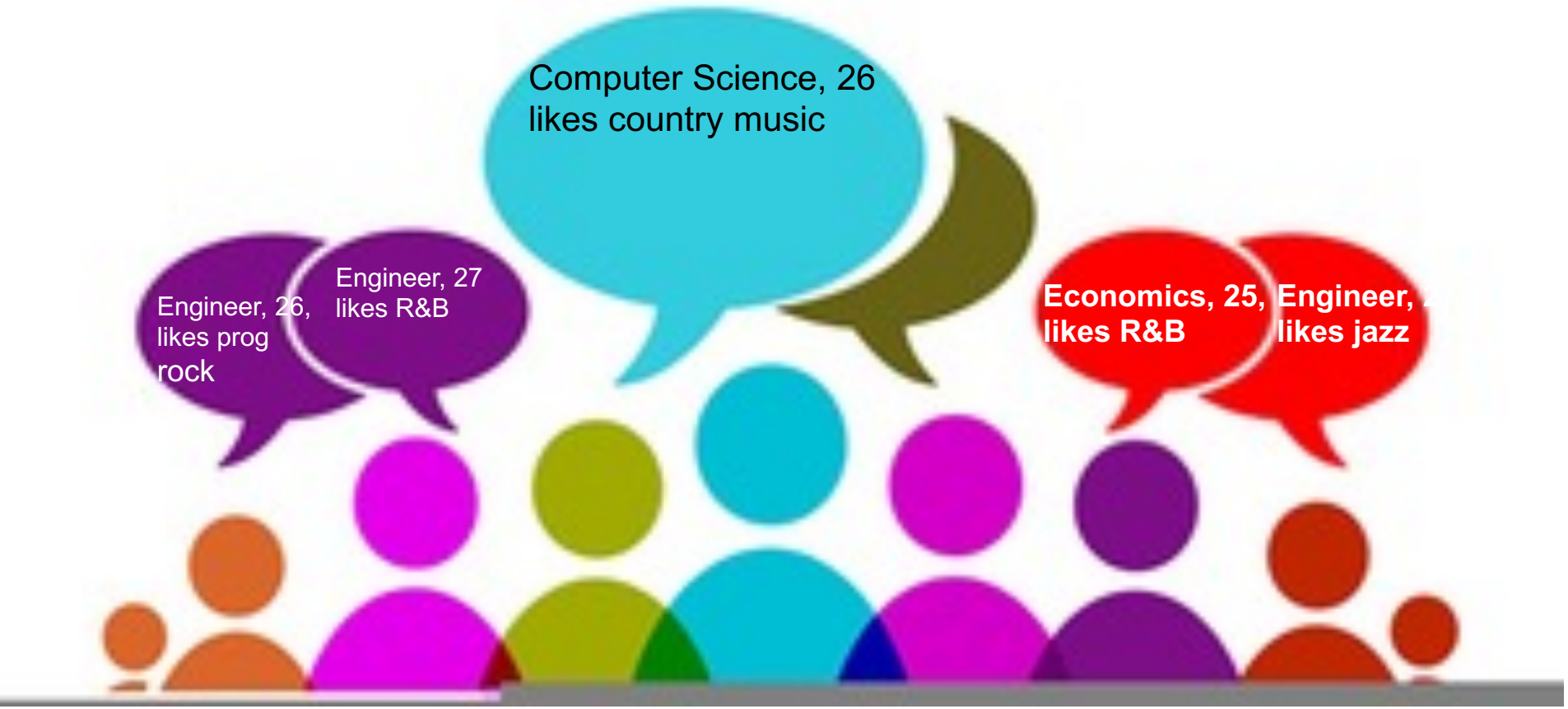


Monthly users of social media (in millions), source: Statista 2022

When Numeric Data Don't Explain Differences in Preferences

Similar demographics, but very different preferences in music

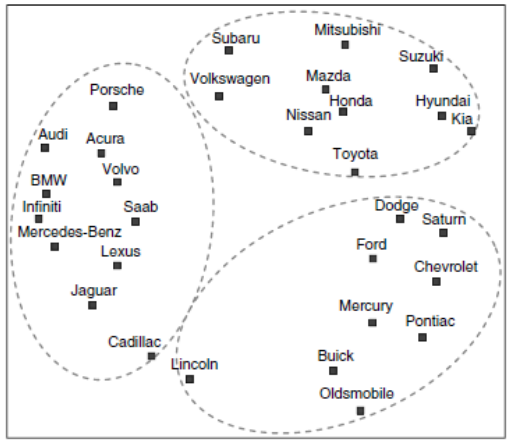
Yet, unstructured data (e.g., blogs) can predict much more accurately



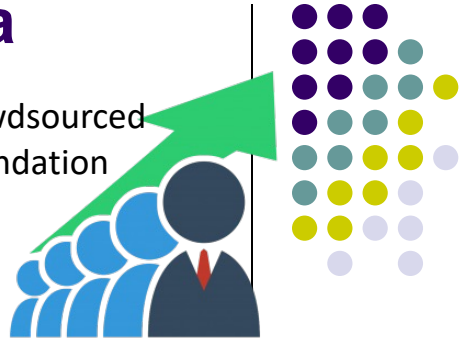
The Untapped Potential of Unstructured Data

Obtain insights: Analyze brand & product associations, predict switching

Well I'm a Benz fan first and foremost, but the Audi 3.0 CVT just offers more
I think both are excellent(sp?) cars, however neither one is a low maintenai
It has come to my attention that as of late MB has taken a hit in its reputatio
There is not much to debate on this subject. If you want a performance car
gave up nationality based generalizations a couple of years ago.
I'm interested more towards performance than luxury. Also value is an imp
cybersol... I can't fathom why you would want to go with slushbox, but that
Sorry to disappoint with the slushbox, but if we were all the same life woul
The new Altima is great as my friends just bought one. However, dollar for
cybersol... Excitement and slushbox don't seem to be compatible. I can't ge
"Excitement and slushbox don't seem to be compatible."Agreed. Unfortuna
Riez, you have heard of performance sensitive steering? This varies the per
cybersol... There are two types. Those tied to performance output and thos
Yeah, I have that brouchure and they don't mention the difference betweer
cybersol... From what I've read, the better variable performance assist syst
So, if you read through most of this group, alot of references went out to Au
Have you driven them both? The S4 really moves. Even by Audi's own 0-60
I've tested the 3.0, not the S4. Hmmm... new S4 in '03, huh? Maybe I should
Yeah, the current S4 has 250 hp and 258 lb-ft of performance. The performa
Not so fast. There is merit to the "nationality based" debate. Let me explai



Build crowdsourced recommendation systems

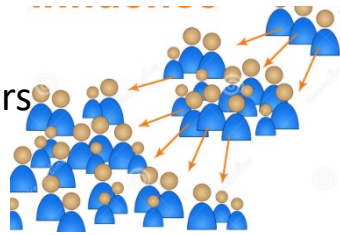


Predict:
E.g., salesrank, retention, spend, etc.

Metrics & analytics

Create new products/services

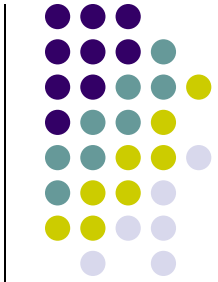
Who matter most: Find influencers
Detect bots, trolls in networks
Analyze real-time events



Before and After this Course

Do you know how to	Yes/No

Final Project (Group Presentations)



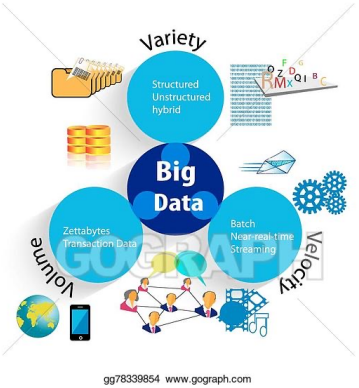
- You are a consultant to a brand, organization, politician, etc.
- Objectives: Obtain actionable insights from social mentions using unstructured data analytics
- Analysis of competing brands or products
  - E.g., airlines, automobiles, hotels, smartphones, etc.
- Track events (e.g., political) or marketing campaigns
  - Whether social mentions reflect the intended messages
- Predict outcomes from social chatter
  - E.g., box office revenues, stock prices, etc.
- Develop a crowdsourced recommendation system
- Use analytics to study the nature of emerging phenomena/movements

Tools, Techniques, ….

- Python scripts (or R, if you prefer, but no help from TA)
- Write your own scripts
  - No coding taught in class (but TA will help if/when required)
  - Scrapers & data access tools (TA will provide demos)
  - GitHub is your friend

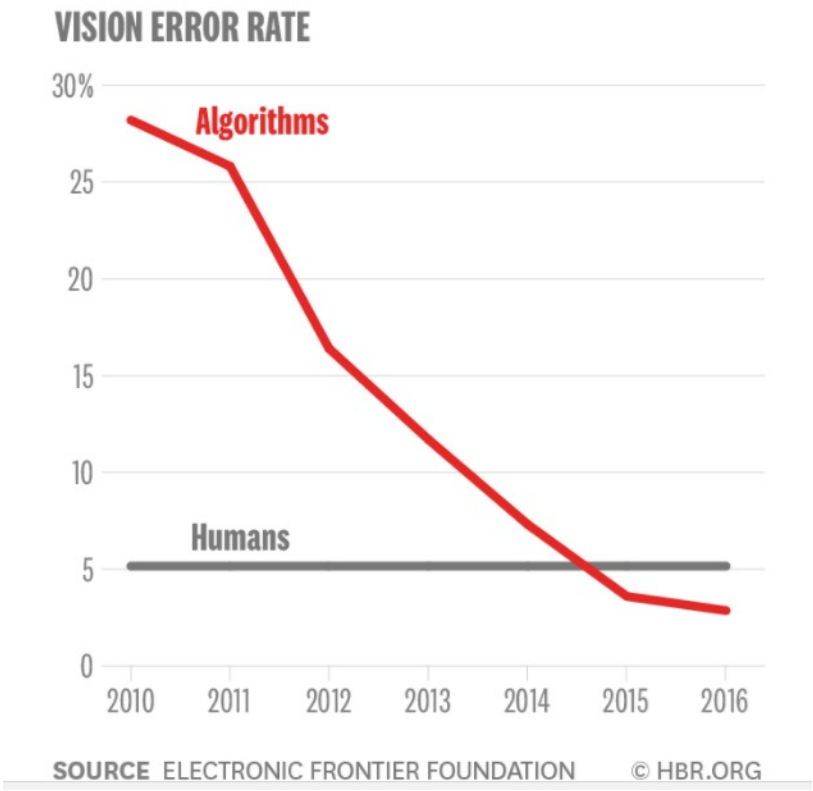
Unstructured Data and Four Vs

What has unstructured data got to do with the Vs?



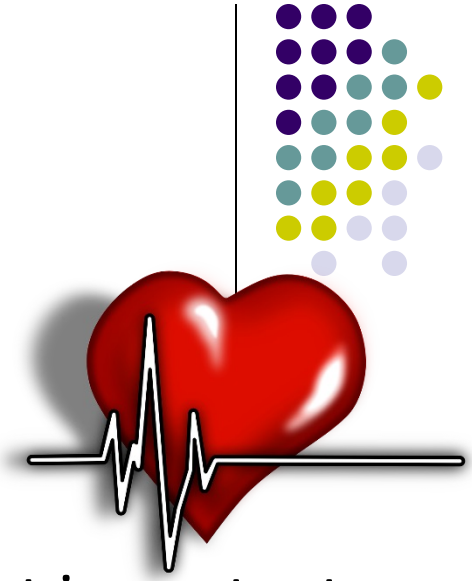
Source: IBM

Image Analytics: Puppy or Muffin?

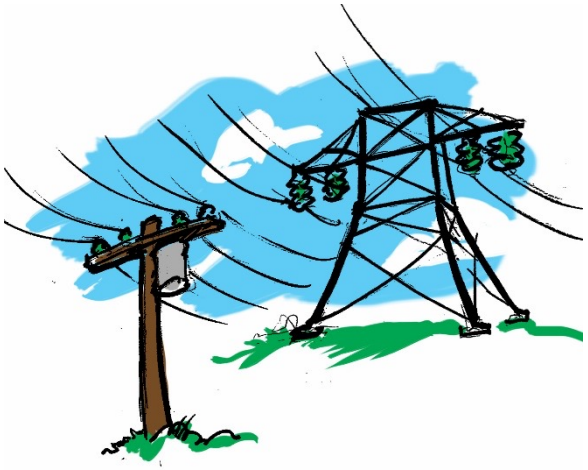


Source: Erik Brynjolfsson

Business Applications With Unstructured Big Data



- Healthcare
  - Which CHF patients will be readmitted?
  - Free-form inputs (text) about lifestyle most important predictor
- Electric grid failures
  - Traditional models
  - Enriching with text



“Told Ya, it’s all Fake!”

When machine learning generates (fake) reviews

1. Easily my favorite Italian restaurant. I love the taster menu, everything is amazing on it. I suggest the carpaccio and the asparagus. Sadly it has become more widely known and becoming difficult to get a reservation for prime times.
2. My family and I are huge fans of this place. The staff is super nice and the food is great. The chicken is very good and the garlic sauce is perfect. Ice cream topped with fruit is delicious too. Highly recommended!
3. I come here every year during Christmas and I absolutely love the pasta! Well worth the price!
4. Excellent pizza, lasagna and some of the best scallops I've had. The dessert was also extensive and fantastic.
5. The food here is freaking amazing, the portions are giant. The cheese bagel was cooked to perfection and well prepared, fresh & delicious! The service was fast. Our favorite spot for sure! We will be back!
6. I have been a customer for about a year and a half and I have nothing but great things to say about this place. I always get the pizza, but the Italian beef was also good and I was impressed. The service was outstanding. The best service I have ever had. Highly recommended.

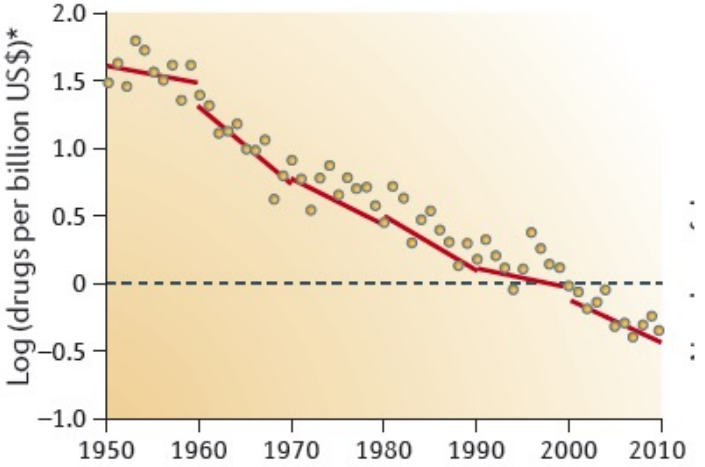
1, 3 & 4 are **real**, **2, 5** and **6** are **fake**

But Machine Learning & Artificial Intelligence can Far Exceed Human Performance

- Harley Davidson dealership in NYC selling 1-2 per week
- Chance encounter with CEO of an AI company
- AI driven marketing platform
- Measures, optimizes campaigns across digital channels
- Sold 15 in first weekend of deployment
- From 1 lead/day → 40, new call center set up
- 15% “lookalikes”
- Month 3: Leads ↑ 2930%, 50% lookalikes
- Redefining “potential customers”
- How does the AI based program do it?

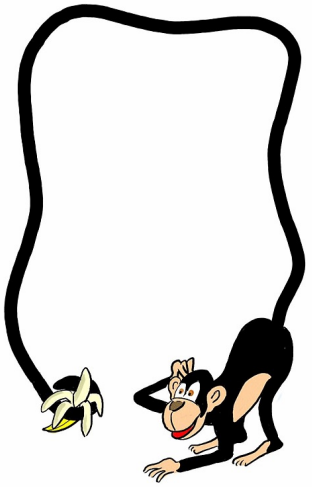
Source: HBSP

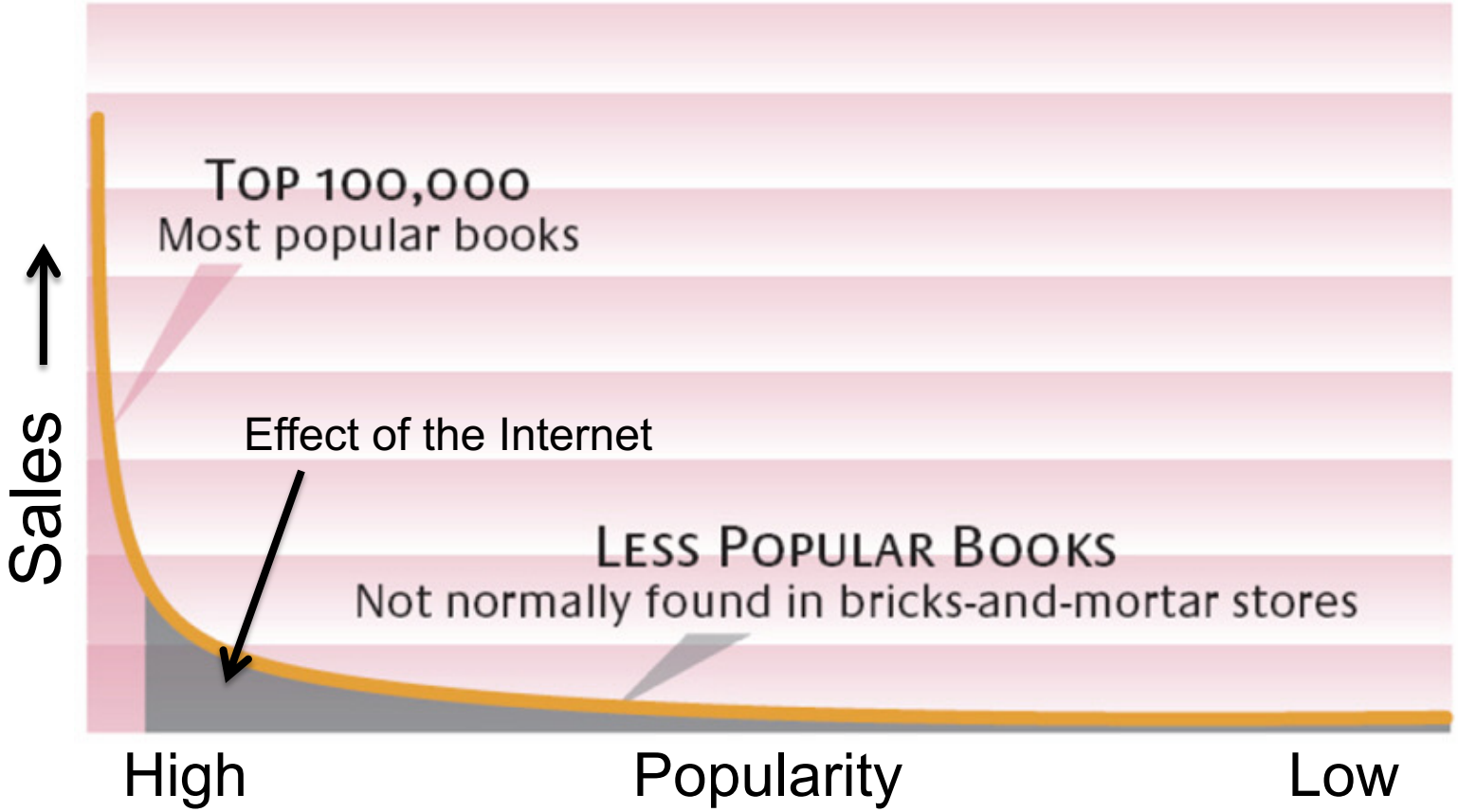
Text Analytics in Pharma: Help Overturn Eroom's Law?



- Biosciences: 10,000 new publications uploaded on a daily basis
- Must correlate, assimilate and connect massive amounts of data
- Make connections between data and generate hypotheses using criteria set by the scientist
- E.g., 200 hypotheses for Amyotrophic Lateral Sclerosis by BenevolentAI

The Connection to Business: The Long Tail*





- Customer preferences are diverse
- What is the role of recommender systems and word-of-mouth?

Source: Brynjolfsson, Hu and Smith; *Chris Anderson

Craft Beers

The Pursuit of Hoppiness

Anita Bhat
Gihani Dissanayake
Alex Jansen
Kyle Katzen
Siddhant Shah

*Students in UT Austin's MSBA program

A “Long Tail” Visibility Problem for Double Sunshine

Double Sunshine IPA | Lawson's Finest Liquids

BA SCORE
4.67/5
2,138 Ratings

BEER STATS

Ranking:	#14
Reviews:	343
Ratings:	2,138
pDev:	7.28%
Bros Score:	0
Wants:	3,276
Gots:	243

Hopslam Ale | Bell's Brewery, Inc.

BA SCORE
4.46/5
13,068 Ratings

BEER STATS

Ranking:	#145
Reviews:	3,532
Ratings:	13,068
pDev:	9.42%
Bros Score:	4.2
Wants:	2,678
Gots:	3,643

A Recommendation for Your Hoppiness

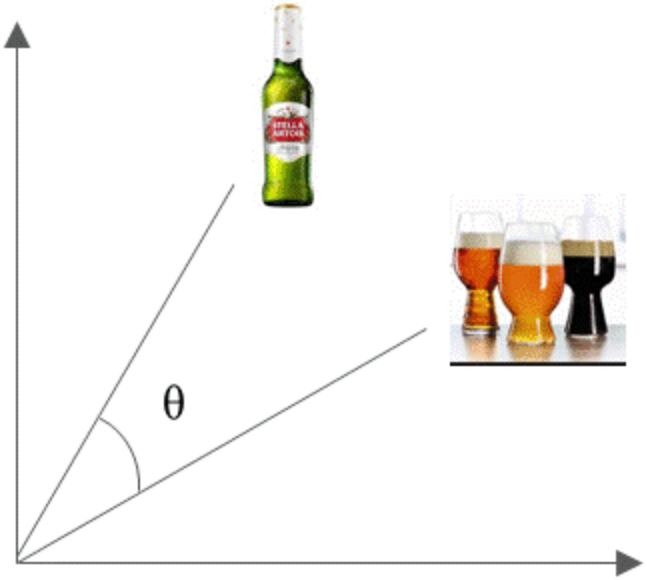
Recommendation Process

Scrape commercial beer reviews

Cosine similarities of commercial beers

Create similarity scores for survey responders

Recommend beers with highest scores



Pros & Cons of Online Social Data

Pros (relative to a survey)	Cons	Mitigation?

Unstructured Data Analytics

Product Preference Networks

MSITM, 31st October, 2022

Dr. Anitesh Barua

David Bruton Jr. Centennial Chair Professor of Business

Distinguished Fellow, INFORMS Information Systems Society

University of Texas Distinguished Teaching Professor

Associate Director, Center for Research in e-Commerce

McCombs School of Business, University of Texas at Austin

Email: aniteshb@gmail.com

Network Analytics Projects: Example #1

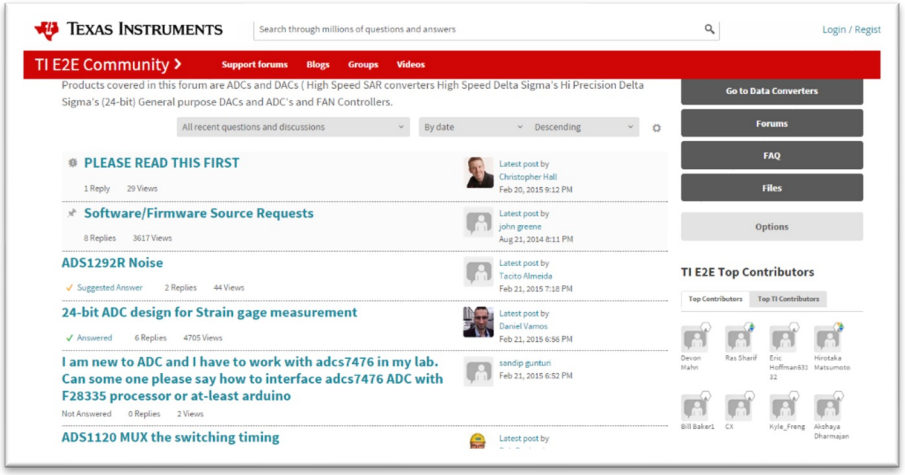
Your Customers Help Each Others or Do They? An Analysis of an Enterprise-to-Enterprise Forum

- The Mantra: "Let customers help each other"
- Cisco and Dell implemented successfully
- Not a new idea, but difficult to succeed with
- Right kind of (typically social) incentives
- Enterprise to Enterprise (E2E) Community of Texas Instruments (TI)
  - Tools and software products
  - Both customers and TI employees
- To what extent are customers helping each other?
- What is the role of TI employees?
- Is TI recognizing the right people?

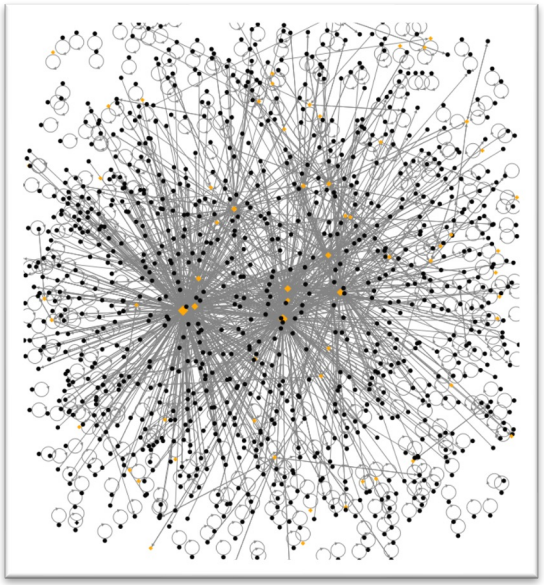
*MSBA project by Megan, Lydia, Anusha, Diana & Tianjiao

The Forum

- http://e2e.ti.com/support/data_converters/precision_data_converters/f/73
  - Python crawler, 24682 messages from one forum
  - Variables: Time, poster, level, points, member type, response to, content (text)
  - What kind of analysis can answer the questions?

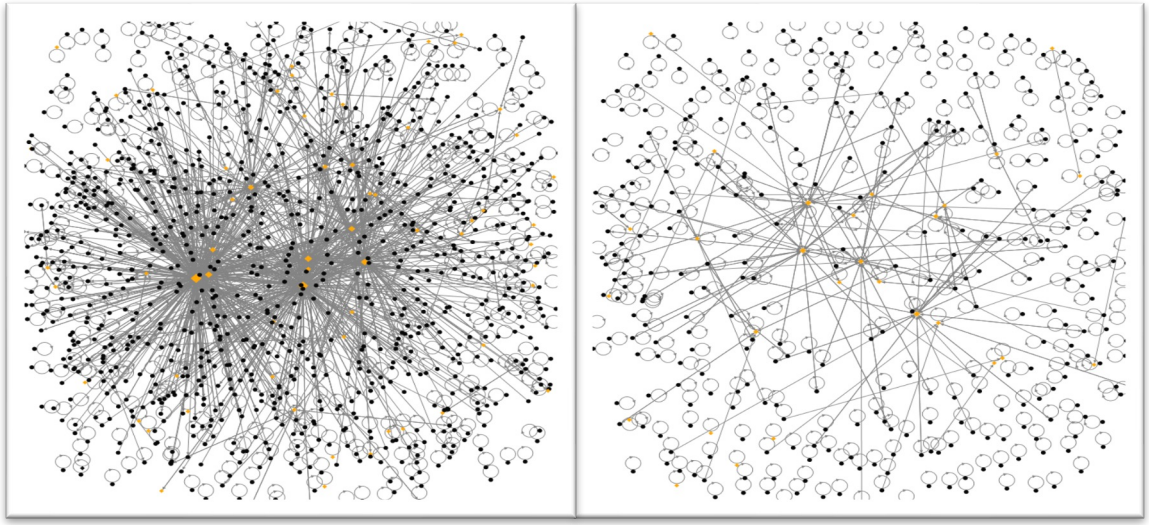


TI Forum Network

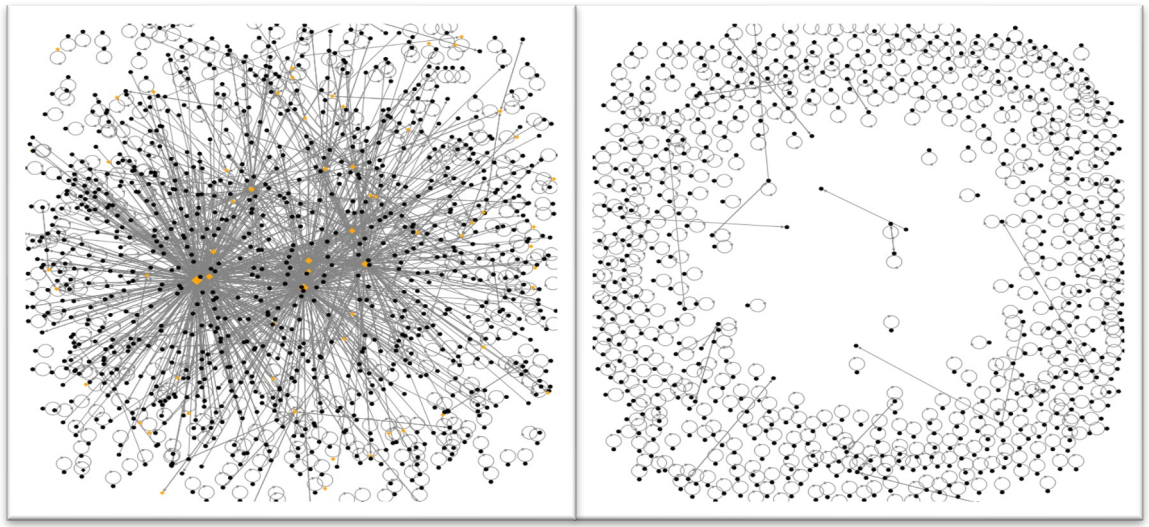


- Includes all participants who have posted ≥ three times.
- TI employees (Orange color) are central
- Many self loops
- Top 20 (by degree) are all TI employees

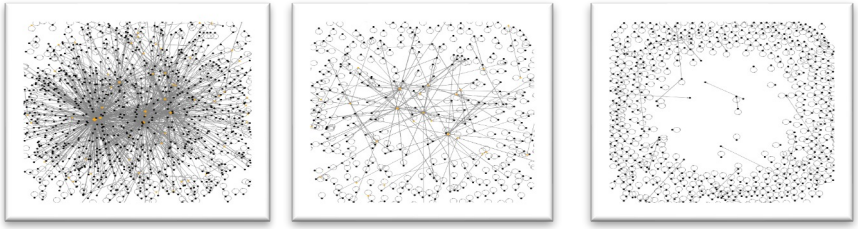
What Happens When we Remove the Top 5?



What Happens if we Remove all TI Employees?



From the Overall Network Perspective



Metrics	Members and TI Employees	Top 5 Missing	No TI Employees
Vertices	1091	828	636
Edges	2092	1130	654
Average Degree	2.8	2.4	2.1
Average Betweeness	1217.501	50.205	0.019
Density	0.0012	0.00066	0.00008

Sentiment Analysis

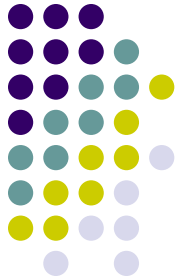
Overall Sentiment

Rank	Name	Average Overall Sentiment	Identity
20		1.615384615	TI Employee
15		1.215686275	TI Employee
14		1.215384615	TI Employee
11		1.198019802	TI Employee
7		1.15862069	TI Employee
13		1.14084507	TI Employee
9		1.1	TI Employee
16		1.1	TI Employee
12		1.094594595	TI Employee
3		1.070844687	TI Employee
1		1.066204288	TI Employee
10		1.048076923	TI Employee
6		1.034161491	TI Employee
4		1.019163763	TI Employee
5		1.016605166	TI Employee
2		1.007686932	TI Employee
17		1	TI Employee
8		0.975694444	TI Employee
18		0.936170213	TI Employee
19		0.225	Community Member

Implications

- Customers are not helping customers!
- Employee performance appraisal
  - Network – visualize the position and calculate the centrality of employees
  - Sentiment Analysis - identify which employees give the highest quality responses (e.g., John Doe)
- Improving TI's technical documents
  - Look at posts with negative sentiment scores
  - Find the posts of "isolated members"

Predicting Business Outcomes From User Generated Content

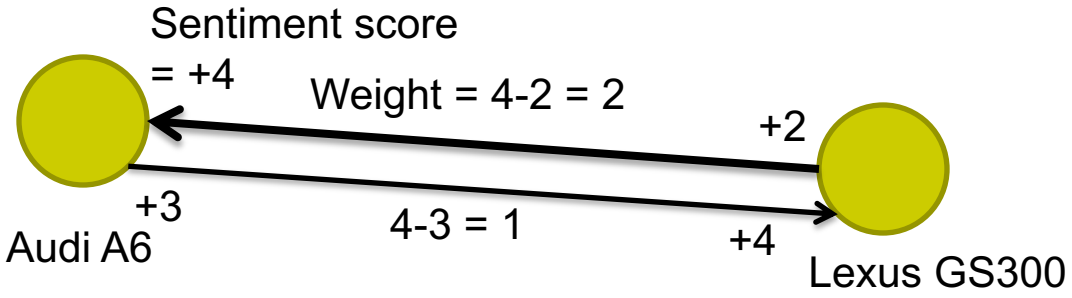


- Can UGC (e.g., product reviews) predict business outcomes such as sales and market share?
- Often users mention competing products in reviews
- Can we extract preference information?
- Can we draw product comparison (or preference) networks?
- Can such networks help predict business outcomes?

Slides based on Zhang et al. (2014)

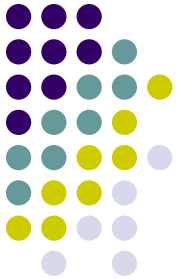
From Product Comparisons to Preference Networks

- "The B&W P7 are high on my favorite list after the H8 by B&O. I also like the new P5 because their sound is almost as good as the P7."
- "I just love the luxury, style and performance of the Audi A6; the Lexus GS300 is a nice reliable car and a very good value, but lacks the coolness factor."



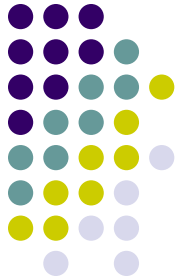
- "I need a super reliable car with the great creature comforts, and while the A6 is a wonderful car, it's quite expensive; the Lexus GS300 isn't exactly the lap of luxury, but really fits the bill for me in every way."

The Main Idea



- Product preference networks
- Arrows indicate implicit preferences
- Relative desirability of a node
- A product review (as discussed in the article)
  - Must mention two or more products
  - Has two sentiment scores (s_1 and s_2) for two products 1 and 2 respectively.
  - Arrow between the two product, tip ends on the product with higher sentiment score
  - Difference in sentiment scores becomes the weight of the arrow
- How to put a score on each node which represents its desirability

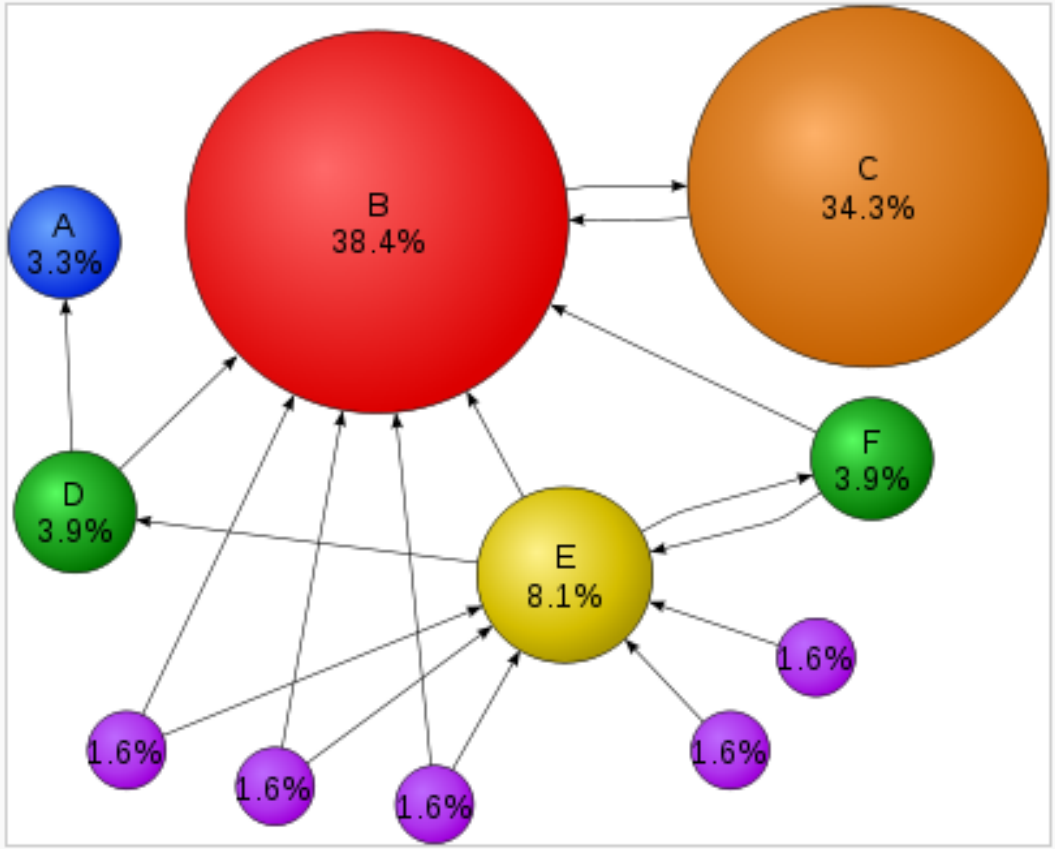
What Metric Can Capture the Relative Importance of a Product?



- Create a network of product preferences
- PageRank is one possibility
- Developed by Larry Page, Serge Brin & Rajiv Motwani at Stanford
- A variation of the good old eigenvector algebra
- Based on how many web pages refer to a particular web page.

PageRank to the Rescue

- If many web pages refer to a page, the latter must be important
- If the referring pages are referred to by many other pages, the effect is stronger



Mathematical **PageRanks** for a simple network, expressed as percentages. (Google uses a logarithmic scale.) Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value. If web surfers who start on a random page have an 85% likelihood of choosing a random link from the page they are currently visiting, and a 15% likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1% of the time. (The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85%.) Without damping, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own.

Source: Wikipedia

The $1T Google Algorithm?

Web page x_1

Links to:
Page x_2
Page x_3
Page x_4

Web page x_3

Links to:
Page x_1

Web page x_2

Links to:
Page x_3
Page x_4

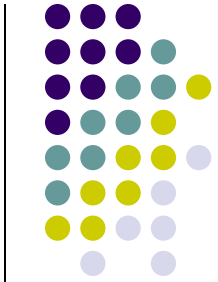
Web page x_4

Links to:
Page x_1
Page x_3

- As if a page transfers its importance to other pages
- E.g., x_1 transfers 1/3 of its importance to each of x_2, x_3 & x_4
- Graph representation, transition matrix, **PageRank** calculation
- Interpretation from a probabilistic perspective

Source: http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html

Problems Galore …



Web page x_1

Links to:
Page x_3

Web page x_3

Links to:
None

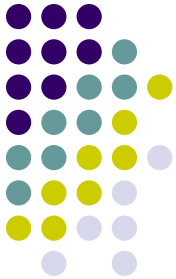
Web page x_2

Links to:
Page x_3

- What does the transition matrix look like?
- Show PageRank Calculations

Source: R. Remus

Problem Cases (Contd.)



Web page x_1

Links to Page x_2

Web page x_3

Links to:
Pages x_4 & x_5

Web page x_5
Links to:
Pages x_3 & x_4

Web page x_2

Links to Page x_1

Web page x_4
Links to:
Page x_3 & x_5

- What did Brin et al. suggest as a solution?

Source: R. Remus

Brin et al.'s Solution

Replace the original transition matrix A by

$M = (1\text{-}p).\ A + p.B$

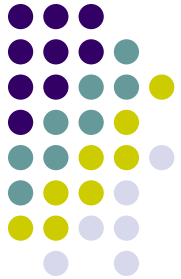
where $0 < p < 1$ (often taken to be a small number, like .1)

$$\text{where } B = \frac{1}{n} \cdot \begin{bmatrix} 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

- What is the interpretation of this version of the transition matrix?
- How does this help?

Source: R. Remus

Weighted PageRank



- Not all incoming links are created equal
- Different ways to create weights on links
- networkx library in python can calculate weighted PageRank scores

Real World Implications