CSE 519: Data Science
Steven Skiena
Stony Brook University

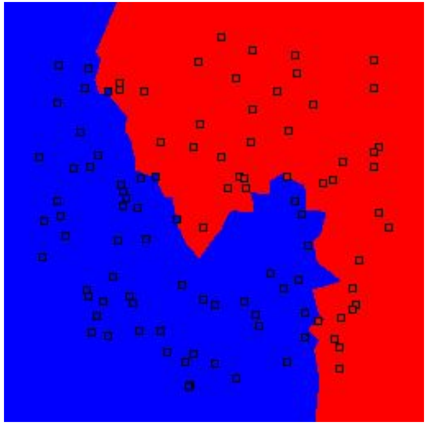
Lecture 19: Nearest Neighbor Methods

Nearest Neighbor Classification

Identify which training example is most similar to the target, and take the class label from it.

The key issue here is devising the right distance function between rows/points.

Advantages: simplicity, interpretability, and non-linearity.



Which Representation is Better?

The nearest neighbors of an item in feature space “should” be like it.

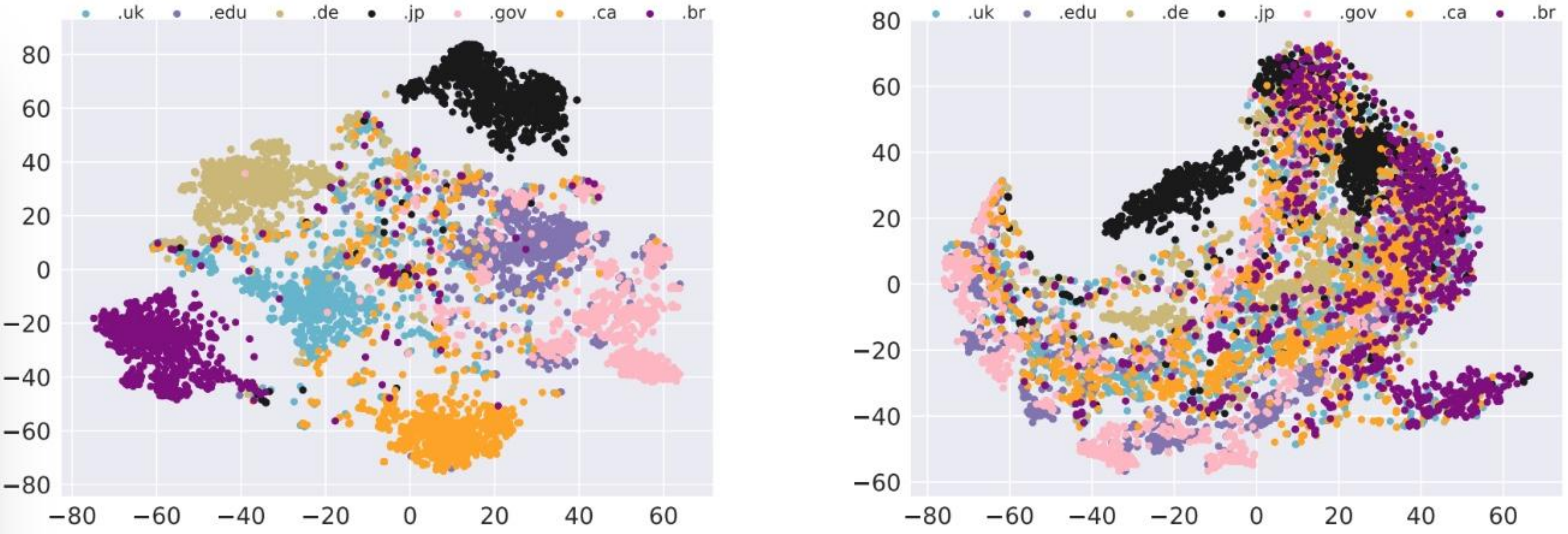
Assessing how well the nearest neighbors of points resemble targets is a great sniff test.

Which of the two feature representations of the web on the next slide is better?

Websites	stonybrook.edu		baidu.com	
Methods	`DeepWalk`	`RandNE`	`DeepWalk`	`RandNE`
Neighbors	binghamton.edu	du.edu	sogou.com	alibaba.com
	www1.cuny.edu	wellesley.edu	qq.com	freepatentsonline.com
	qcpages.qc.cuny.edu	illinois.edu	youku.com	wolframalpha.com
	lehman.edu	amherst.edu	163.com	wikimediafoundation.org
	barnard.edu	ohio.edu	renren.com	news.softpedia.com
	esf.edu	binghamton.edu	tudou.com	worldwide.espacenet.com
	baruch.cuny.edu	smith.edu	naver.com	duckduckgo.com
	colgate.edu	vanderbilt.edu	taobao.com	whois.domaintools.com
	hunter.cuny.edu	macalester.edu	t.qq.com	images.apple.com
	hamilton.edu	lsa.umich.edu	baike.baidu.com	tineye.com

Websites	chase.com		wikipedia.org	
Methods	`DeepWalk`	`RandNE`	`DeepWalk`	`RandNE`
Neighbors	bankofamerica.com	fedex.com	wikipedia.com	timeanddate.com
	capitalone.com	travelocity.com	wikimediafoundation.org	groups.google.com
	citi.com	priceline.com	wikimedia.org	wolframalpha.com
	schwab.com	bankofamerica.com	openoffice.org	digg.com
	wellsfargo.com	capitalone.com	addons.mozillar.org	alexa.com
	discovercard.com	comcast.com	answers.com	spreadsheets.google.com
	creditcards.com	jdpower.com	wolframalpha.com	earth.google.com
	ameriprise.com	delta.com	dmoz.org	xkcd.com
	mastercard.us	ups.com	en.wiktionary.org	quora.com

Clusters by High Level Domain



(a) DeepWalk.   (b) RandNE.

Distance Metrics

Certain mathematical properties are expected of any distance measure, or *metric*:

- $d(x,y)>=0$ for all x, y (positivity)
- $d(x,y)=0$ iff $x=y$ (identity)
- $d(x,y)=d(y,x)$ (symmetry)
- $d(x,y)<=d(x,z)+d(z,y)$ (triangle inequality)

Not a Metric

Many natural similarity measures are not distance metrics:

- Correlation coefficient (-1 to 1)
- Cosine similarity / dot product (-1 to 1)
- Mutual information measures (often not symmetric)
- Cheapest airfare (think triangle inequality)

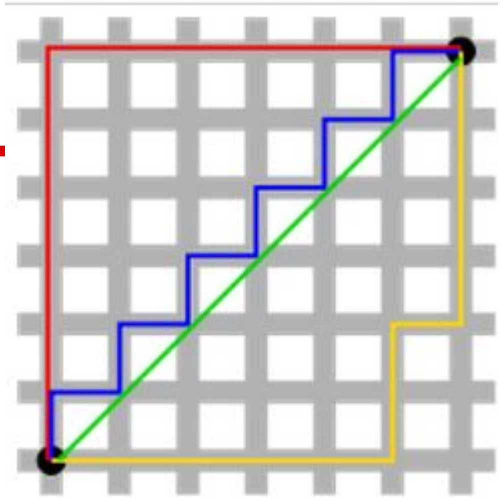
Euclidean Distance Metric

The traditional Euclidean distance metric weighs all dimensions equally:

$$\mathrm{d(x,y)} = \sqrt{\sum_{i=1}^{d} |x_i - y_i|^2}$$

We might use a coefficient c_i to give a different weight to each dimension, but *at least* normalize to make dimensions comparable.

L_k Distance Norms



To generalize Euclidean distance:

$$d(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^k\right)^{(1/k)}$$

- k=1 gives the Manhattan distance metric
- $k = \infty$ gives the maximum component

k regulates the trade off between largest and total dimensional difference.

Which Point is Farther from (0,0)?

Is p1=(2,0) or p2=(1.5,1.5) farther from (0,0)?

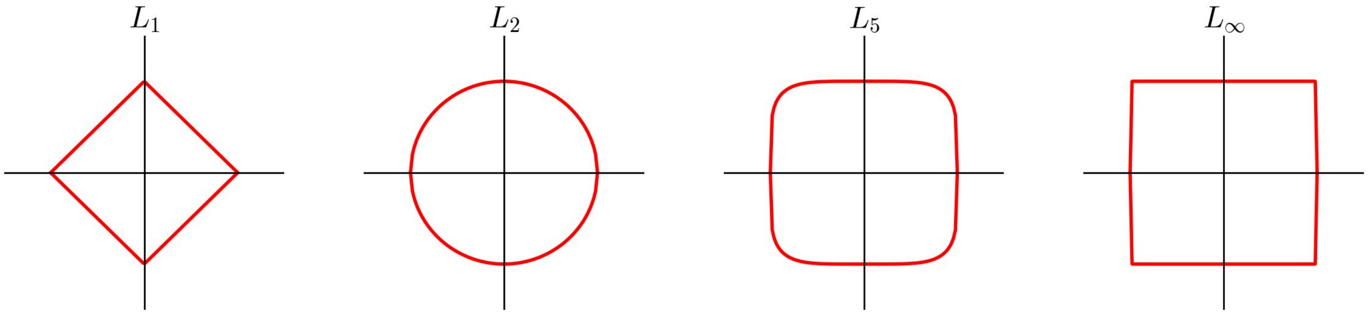
- For k=1, the distances are 2 and 3, so p2
- For k=2, the distances are 2 and 2.12, so p2
- For k=infty, distances are 2 and 1.5, so p1

The distance metric sets which point is closer.

Are we more worried about random noise or dimensional outliers/artifacts?

Circles for Different *k*

The shape of the L_k "circle" governs which points are equal neighbors about the origin.



The distinction here become particularly important in higher dimensional spaces: do we care about deviations in all dimensions or primarily the biggest?

Projections from Higher Dimensions

Projection methods (like SVD) compress or ignore dimensions to reduce representation complexity.

Nearest neighbors in such spaces can be more robust than in the original space.

Dimensional Egalitarianism

Although L_k norms in principle weigh all dimensions equally, the scale matters.

But note that the real impact of height will differ depending upon whether it is measured in centimetres, meters, or kilometers.

This is why we use Z-scores for normalization!

Regression / Interpolation by NN

The idea of nearest neighbor classification can be generalized to function interpolation, by averaging the values of the k nearest points.

Weighted averaging schemes can value points differently according to (1) distance rank, (2) actual distances.

Similar ideas work for all classification methods.

K-Nearest Neighbors

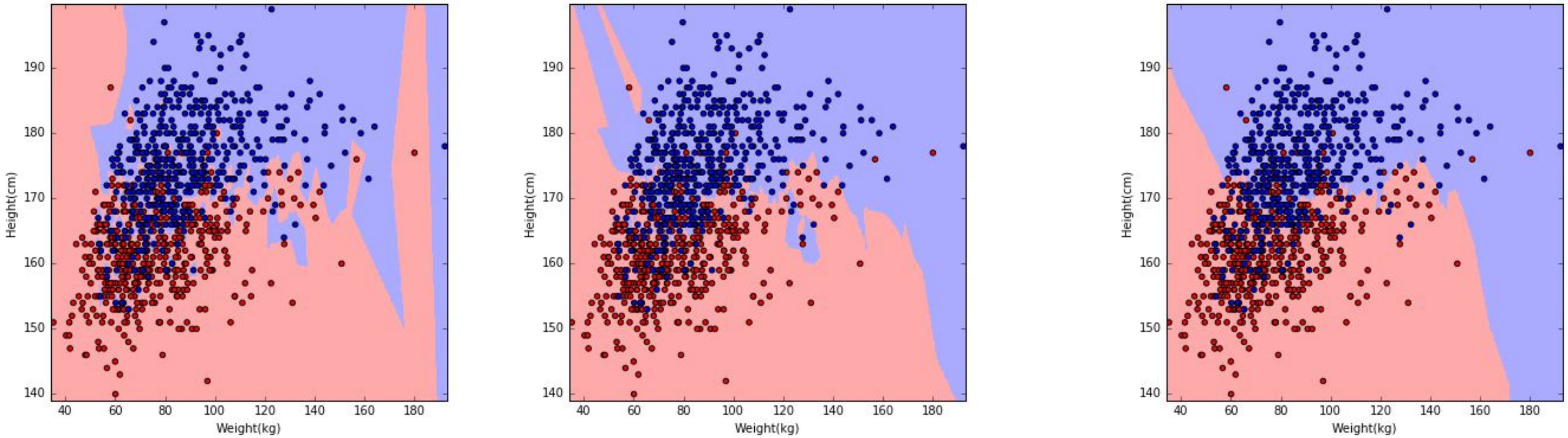
NN classification produce non-linear classifiers, because each training point changes the separating boundary.

More robust classification or interpolation follow from voting over the *k* closest neighbors for *k*>*1*.

Gender Classification by Height/Weight

Smoother boundaries follow from larger k:

k=1 k=3 k=10



Seeking Good Analogies

Many intellectual disciplines rest on analogies:

- Law: which legal precedent was most like this case?
- Medicine: how did I treat patients with similar symptoms, and did they survive?
- Real estate: what price did comparable properties sell for in the neighborhood?

Nearest Neighbors: Mike Trout

Similarity Scores

Explanation of Similarity Scores

Similar Batters

1. Hack Wilson (880.8) *
2. Giancarlo Stanton (856.4)
3. J.D. Drew (856.1)
4. Paul Goldschmidt (849.2)
5. Wally Berger (848.6)
6. Carlos Gonzalez (840.3)
7. Larry Doby (839.2) *
8. Ralph Kiner (835.8) *
9. Eric Davis (833.3)
10. Darryl Strawberry (832.9)

* - Signifies Hall of Famer

Similar Batters through 28

1. Frank Robinson (922.3) *
2. Mickey Mantle (901.7) *
3. Ken Griffey Jr. (889.3) *
4. Duke Snider (883.8) *
5. Hank Aaron (878.9) *
6. Willie Mays (869.5) *
7. Eddie Mathews (867.0) *
8. Juan Gonzalez (861.9)
9. Andruw Jones (858.7)
10. Miguel Cabrera (853.3)

* - Signifies Hall of Famer

Most Similar by Ages

20. Vada Pinson (955.2) 2 3 4 5 6 7 8 9 10 C
21. Frank Robinson (957.8) * 2 3 4 5 6 7 8 9 10 C
22. Mickey Mantle (941.3) * 2 3 4 5 6 7 8 9 10 C
23. Mickey Mantle (941.1) * 2 3 4 5 6 7 8 9 10 C
24. Mickey Mantle (960.6) * 2 3 4 5 6 7 8 9 10 C
25. Frank Robinson (948.8) * 2 3 4 5 6 7 8 9 10 C
26. Frank Robinson (955.4) * 2 3 4 5 6 7 8 9 10 C
27. Mickey Mantle (935.8) * 2 3 4 5 6 7 8 9 10 C
28. Frank Robinson (922.3) * 2 3 4 5 6 7 8 9 10 C

* - Signifies Hall of Famer

Seeking Good Analogies? (Projects)

- Miss Universe?
- Movie gross?
- Baby weight?
- Art auction price?
- Snow on Christmas?
- Super Bowl / College Champion?
- Ghoul Pool?
- Future Gold / Oil Price?

Finding Nearest Neighbors

Given *n* points in *d*-dimensions, it takes $O(nd)$ time to find the NN using brute force search.

For large training sets or high dimensionality this becomes very expensive.

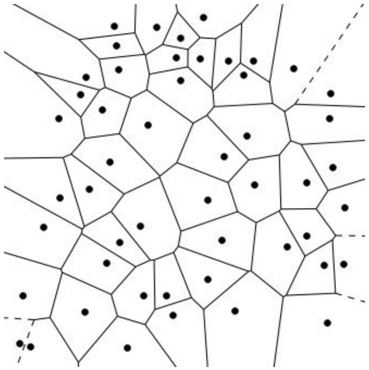
This motivates use of more sophisticated data structures: grid indices, kd-trees, Voronoi diagrams, and locality sensitive hashing.

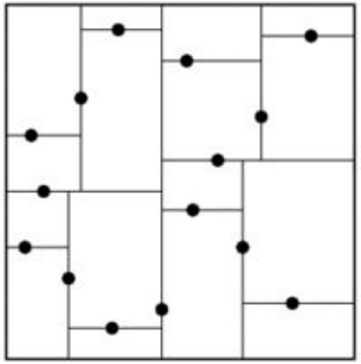
Voronoi Diagrams / Kd-trees

Voronoi diagrams partition space into regions sharing nearest neighbors.

Efficient algorithms exist for finding nearest neighbors in low dimensions, such as kd-trees and ball trees.

But exact NN search is doomed to reduce to linear search for high-enough dimensionality data.

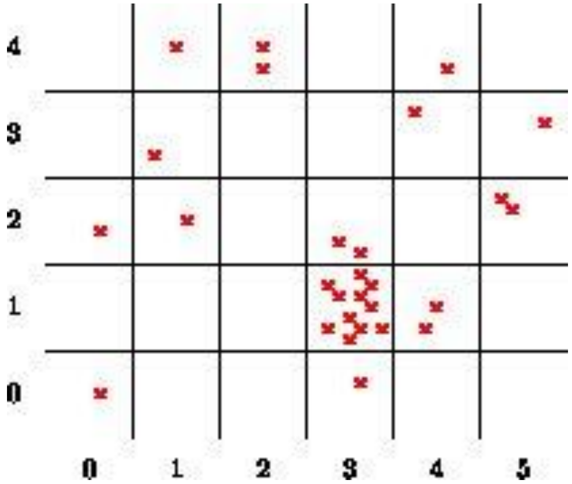




Grid Files

Bucketting points on a regularly-spaced grid provides a way to group points by similarity.

But such an index becomes expensive as the number of dimensions increases.



Locality Sensitive Hashing

Hashing could speed nearest-neighbor search if nearby points got hashed to the same bucket.

But normal hashing uses hash functions that spreads similar items to distance buckets.

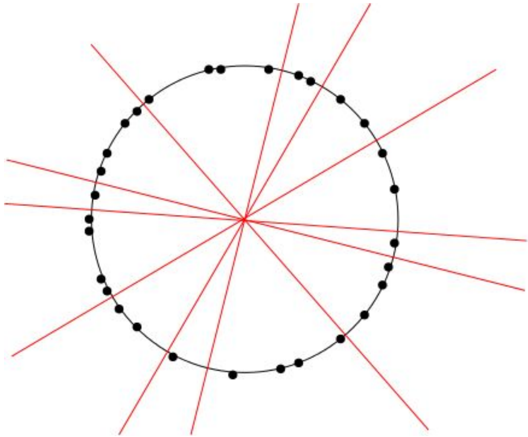
Locality sensitive hashing (LSH) takes points or vectors a and b such that likely h(a)=h(b) iff a is near b.

LSH for Points on a Sphere

Pick random planes cutting through the origin.

If near each other, two points are likely on the same side (left or right) of a given random plane.

L/R patterns for d random planes define a d-bit LSH hash code.



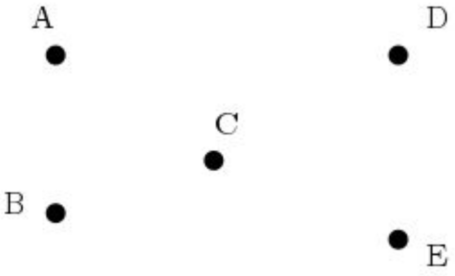
Network Data

Many datasets have natural interpretations as graphs/networks:

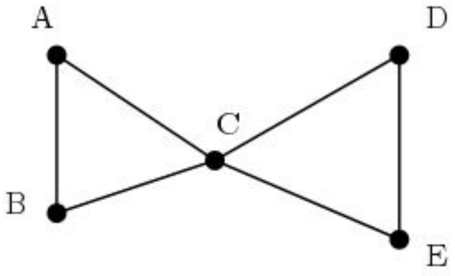
- Social networks: vertices are people, edges are friendships.
- WWW: vertices=pages, edges=hyperlinks.
- Product/customer networks: edges=sales.
- Genetic networks: v=genes, e=interactions.

Point Sets and Graphs

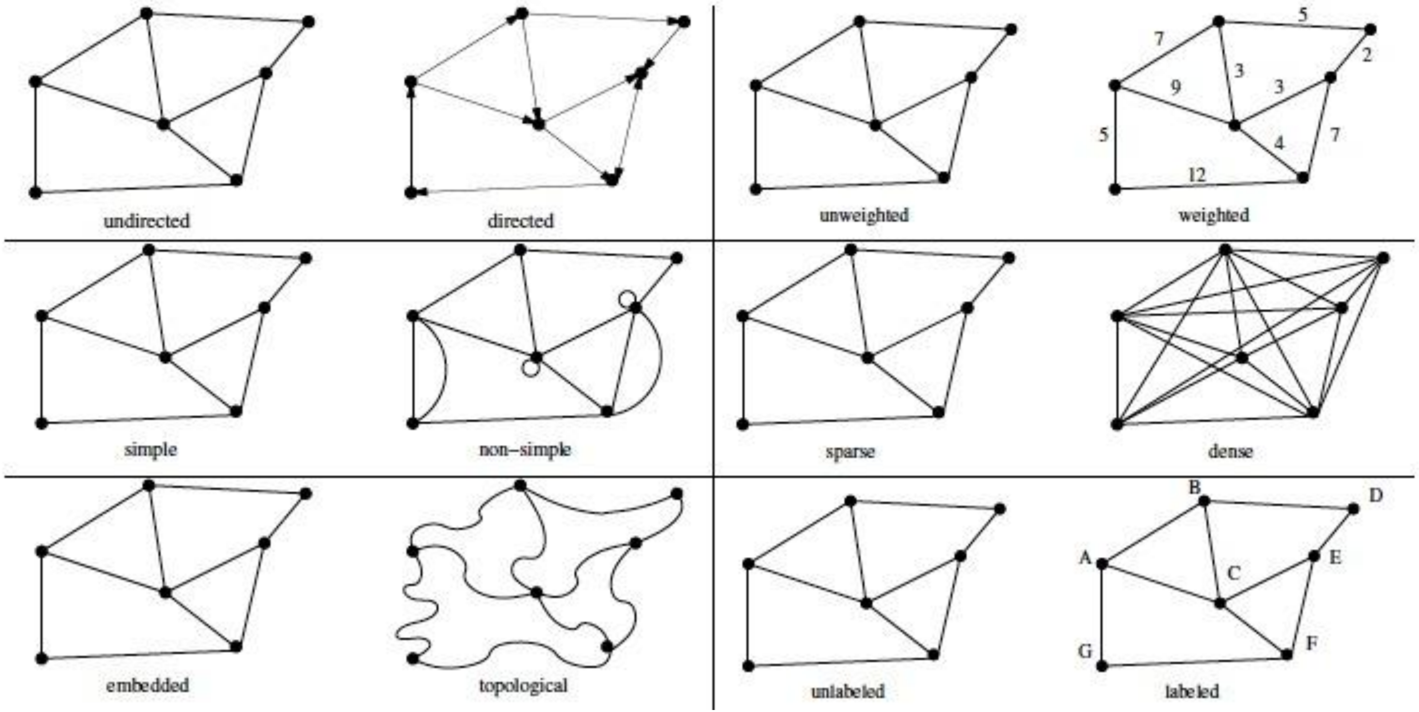
- Point sets naturally define graphs: add an edge *(x,y)* if *x* and *y* are close enough.
- Graphs naturally define point sets: perform an SVD of the adjacency matrix.



	A	B	C	D	E
A	0	6	7.2	13	14.8
B	6	0	6.3	14.3	13.0
C	7.2	6.3	0	8.1	7.6
D	13	14.3	8.1	0	4
E	14.8	13.0	7.6	4	0



Taxonomies of Graph Types



Talking the talk is part of walking the walk.

Classical Graph Algorithms

Distances in graphs are naturally defined in terms of the shortest path between vertices.

Classical algorithms for finding shortest paths, connected components, spanning trees, cuts, flows, matchings, topological sorting can be applied to any appropriate network.

PageRank

PageRank(v,G) corresponds to the probability that a random walk on *G* ends up at vertex *v*.

The basic formula is:

$$PR_j(v) = \sum_{(u,v)\in E} \frac{PR_{j-1}(u)}{out-degree(u)}$$

This recursive formula defines an iterative algorithm, which quickly converges in practice.

If All Roads Lead to Rome...

… then Rome must be an important place.

PageRank is a good feature to measure centrality or importance of vertices, like Wikipedia pages:

PageRank PR1 (all pages)	
1	Napoleon
2	George W. Bush
3	Carl Linnaeus
4	Jesus
5	Barack Obama
6	Aristotle
7	William Shakespeare
8	Elizabeth II
9	Adolf Hitler
10	Bill Clinton

PageRank PR2 (only people)	
1	George W. Bush
2	Bill Clinton
3	William Shakespeare
4	Ronald Reagan
5	Adolf Hitler
6	Barack Obama
7	Napoleon
8	Richard Nixon
9	Franklin D. Roosevelt
10	Elizabeth II

PageRank in Practice

Design decisions in PageRank include:

- Editing the graph to remove irrelevant vertices/edges (spam).
- Removing outdegree zero vertices.
- Allowing random jumps: the damping factor.

PageRank is less important to Google today than popularly supposed.

Graph Embeddings (DeepWalk)

Networks based on similarity or links define very sparse feature vectors.   Techniques like SVD can compress the matrix, but are expensive.

Pairs of vertices which often appear near each other on random walks on networks.

Thus we can use data from random walk neighborhoods to train network representations!

Nearest Neighbors in Wikipedia

The links between pages defines the network.

Ludwig van Beethoven
- Franz Schubert (0.489)
- Johannes Brahms (0.532)
- Wolfgang Mozart (0.567)
- Robert Schumann (0.576)
- Gustav Mahler (0.635)

Barack Obama
- George W. Bush (0.474)
- Hillary Clinton (0.657)
- Bill Clinton (0.658)
- Joe Biden (0.750)
- Al Gore (0.791)

Scarlett Johansson
- Kirsten Dunst (0.784)
- Natalie Portman (0.786)
- Gwyneth Paltrow (0.796)
- Brad Pitt (0.858)
- Cameron Diaz (0.891)

Mick Jagger
- John Lennon (0.687)
- Keith Richards (0.687)
- Paul McCartney (0.796)
- Ronnie Wood (0.822)
- Eric Clapton (0.833)

Albert Einstein
- Richard Feynman (1.049)
- Max Planck (1.073)
- Freeman Dyson (1.107)
- Stephen Hawking (1.153)
- Robert Oppenheimer (1.156)

Steven Skiena
- Larry Page (1.597)
- Sergey Brin (1.598)
- Danny Hillis (1.644)
- Andrei Broder (1.652)
- Mark Weiser (1.653)

Random Projection Graph Embeddings

Random projection methods compute the dot product of each adjacency matrix row with random vectors.

Similar rows should produce similar dot products:

- [0,1,1,0,1,1,0] [5,-2,1,4,2,-3,-1] = -2+1+2-3 = -2
- [0,1,0,0,1,1,0] [5,-2,1,4,2,-3,-1] = -2+2-3 = -1
- [1,0,1,1,0,0,1] [5,-2,1,4,2,-3,-1] = 5+1+4-2 = 8

The Johnson-Lindenstrauss theorem says random projection in (log n)^2 dimensions preserves distances.