



Computer Science in the Information Age

John Hopcroft
Cornell University
Ithaca, New York



Time of change

- There is a fundamental revolution taking place that is changing all aspects of our lives.
- Those individuals who recognize this and position themselves for the future will benefit enormously.



Early years of Computer Science

- Programming languages
- Compilers
- Operating systems
- Network protocols
- Algorithms
- Computability



Computer Science is changing

- Structure of large networks
- Large data sets
- Information
- Search




Drivers of change

- Computers are becoming ubiquitous
- Speed sufficient for word processing, email, chat and spreadsheets
- Merging of computing and communications
- Data available in digital form
- Devices being networked



Computer Science departments are beginning to develop courses that cover the underlying theory

- Random graphs
- Phase transitions
- Giant components
- Spectral analysis
- Small world phenomena
- Grown graphs



What is the theory needed to support the future?

- Large amounts of data
- Noisy data with outliers
- High dimensional
- Possibly power law distributed



Internet queries

Today

- Autos
- Graph theory
- Colleges, universities
- Computer science

Internet queries are changing

Today

- Autos
- Graph theory

- Colleges, universities
- Computer science

Tomorrow

- Which car should I buy?
- Construct an annotated bibliography on graph theory
- Where should I go to college?
- How did the field of CS develop?

What car should I buy?

- List of makes
 - Cost
 - Reliability
 - Fuel economy
 - Crash safety
- Pertinent articles
 - Consumer reports
 - Car and driver

Where should I go to college?

- List of factors that might influence choice

- Cost
- Geographic location
- Size
- Type of institution

- Metrics

- Ranking of programs
- Student faculty ratios
- Graduates from your high school/neighborhood

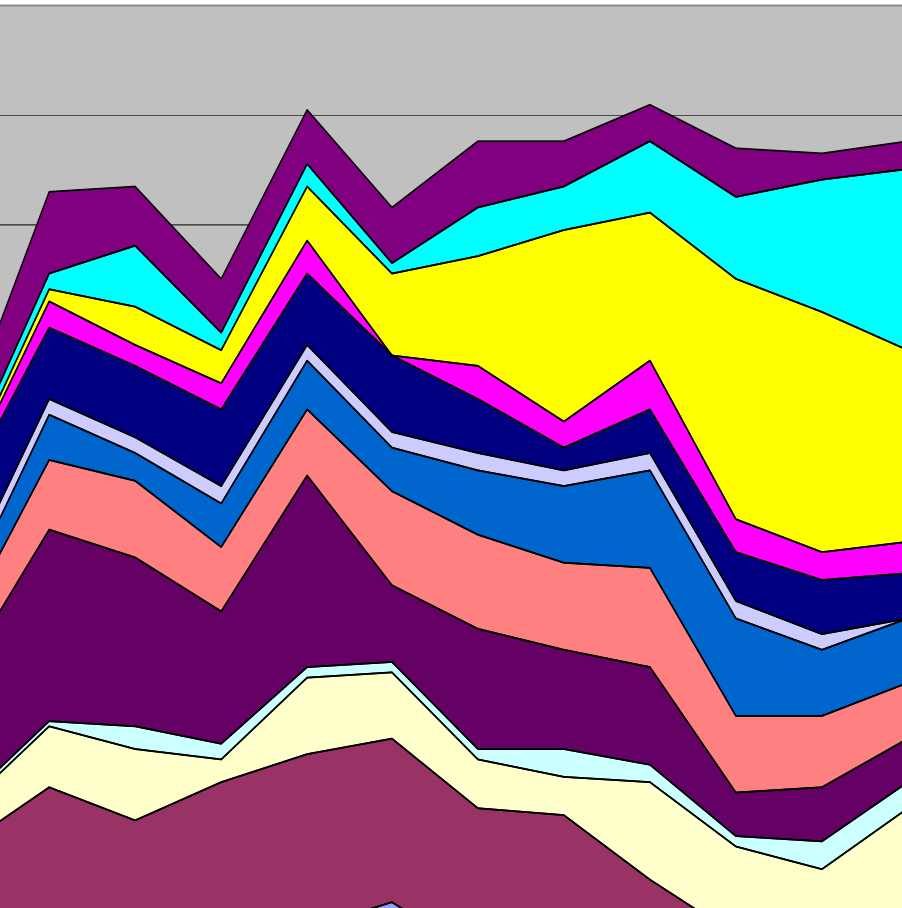
How did field develop?

- From ISI database create set of important papers in field?
- Extract important authors
 - author of key paper
 - author of several important papers
 - thesis advisor of Ph.D.'s student(s) who is (are) important author(s)
- Extract key institutions
 - institution of important author(s)
 - Ph.D. granting institution of important authors
- Output
 - Directed graph of key papers
 - Flow of Ph.D.'s geographically with time
 - Flow of ideas

Topical Cluster Histograms: NIPS Results



NIPS k-means clusters (k=13)



LAS VEGAS, NV • 9/28-29

- 12: chip, circuit, analog, voltage, vlsi
- 11: kernel, margin, svm, vc, xi
- 10: bayesian, mixture, posterior, likelihood, em
- 9: spike, spikes, firing, neuron, neurons
- 8: neurons, neuron, synaptic, memory, firing
- 7: david, michael, john, richard, chair
- 6: policy, reinforcement, action, state, agent
- 5: visual, eye, cells, motion, orientation
- 4: units, node, training, nodes, tree
- 3: code, codes, decoding, message, hint
- 2: image, images, object, face, video
- 1: recurrent, hidden, training, units, error

Publishing

- Researcher makes discovery and writes technical paper
- Submits paper to journal
- Journal sends it out for refereeing
- Revised article is copy edited and appears in print about two years later
- Results are available world wide through major research libraries



The future of publishing

- Advice to young faculty forced journals to let Google search them and ultimately to allow authors to post their article on their website.
- What other changes are in store?

Wikipedia

- 905,707 articles
- Recent comparison showed Wikipedia almost as accurate as Encyclopedia Britannica
- Major text source for formulas, definitions and proofs

Cayley's formula

From Wikipedia, the free encyclopedia.

Jump to: [navigation](#), [search](#)

In [mathematics](#), **Cayley's formula** is a result in [graph theory](#). It states that if n is a positive integer, the number of [trees](#) on n labeled [vertices](#) is

$$n^{n-2}.$$

It is a particular case of [Kirchhoff's theorem](#).

[\[edit\]](#)

Proof of the formula

Let T_n be the set of trees possible on the vertex set $\{v_1, v_2, \dots, v_n\}$. We seek to show that $|T_n| = n^{n-2}$.

We begin by proving a lemma:

Claim: Let d_1, d_2, \dots, d_n be positive integers such that $\sum_{i=1}^n d_i = 2n - 2$. Let \mathcal{A} be the set of trees on the vertex set $\{v_1, v_2, \dots, v_n\}$ such that the degree of v_i (denoted $d(v_i)$) is d_i for $i = 1, 2, \dots, n$. Then

$$|\mathcal{A}| = \frac{(n-2)!}{(d_1-1)!(d_2-1)! \cdots (d_n-1)!}.$$

Proof: We go by induction on n . For $n = 1$ and $n = 2$ the proposition holds trivially and is easy to verify. We move to the inductive step. Assume $n > 2$ and that the claim holds for degree sequences on $n - 1$ vertices. Since the d_i are all positive but their sum is less than $2n$,

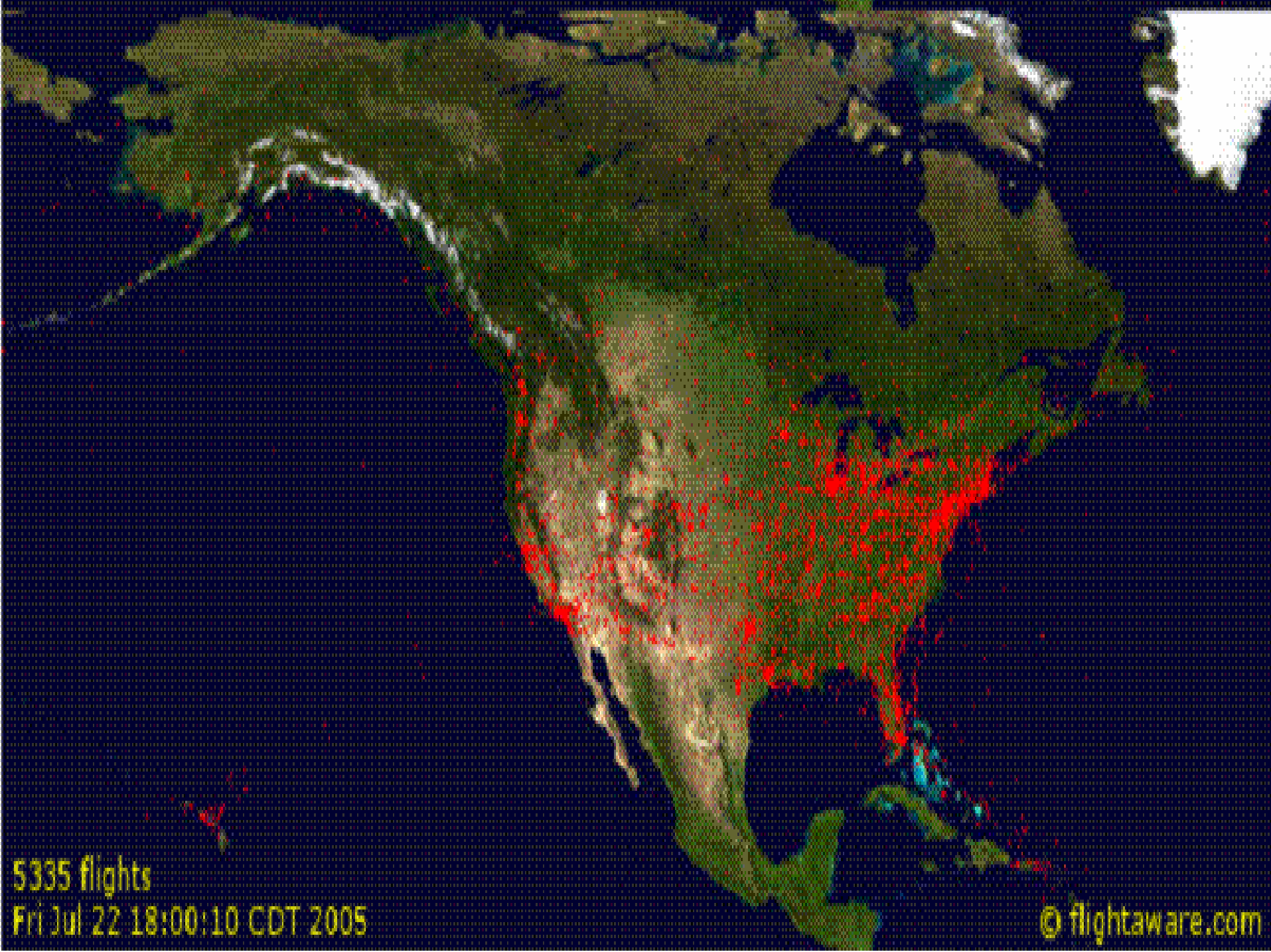
$\exists k \in \{1, 2, \dots, n\}$ such that $d_k = 1$. Assume without loss of generality that $d_n = 1$.

For $i = 1, 2, \dots, n - 1$ let \mathcal{B}_i be the set of trees on the vertex set $\{v_1, v_2, \dots, v_{n-1}\}$ such that:

Fed Ex package tracking

Tracking number	XXXXXXXXXXXXXXXXXX
Ship date	Dec 16, 2005
Delivered to	Receptionist/Front Desk
Destination	Ithaca, NY
Delivery date	Dec 19, 2005 9:28 AM
Signed for by	J. SMITH
Service type	Priority Pak

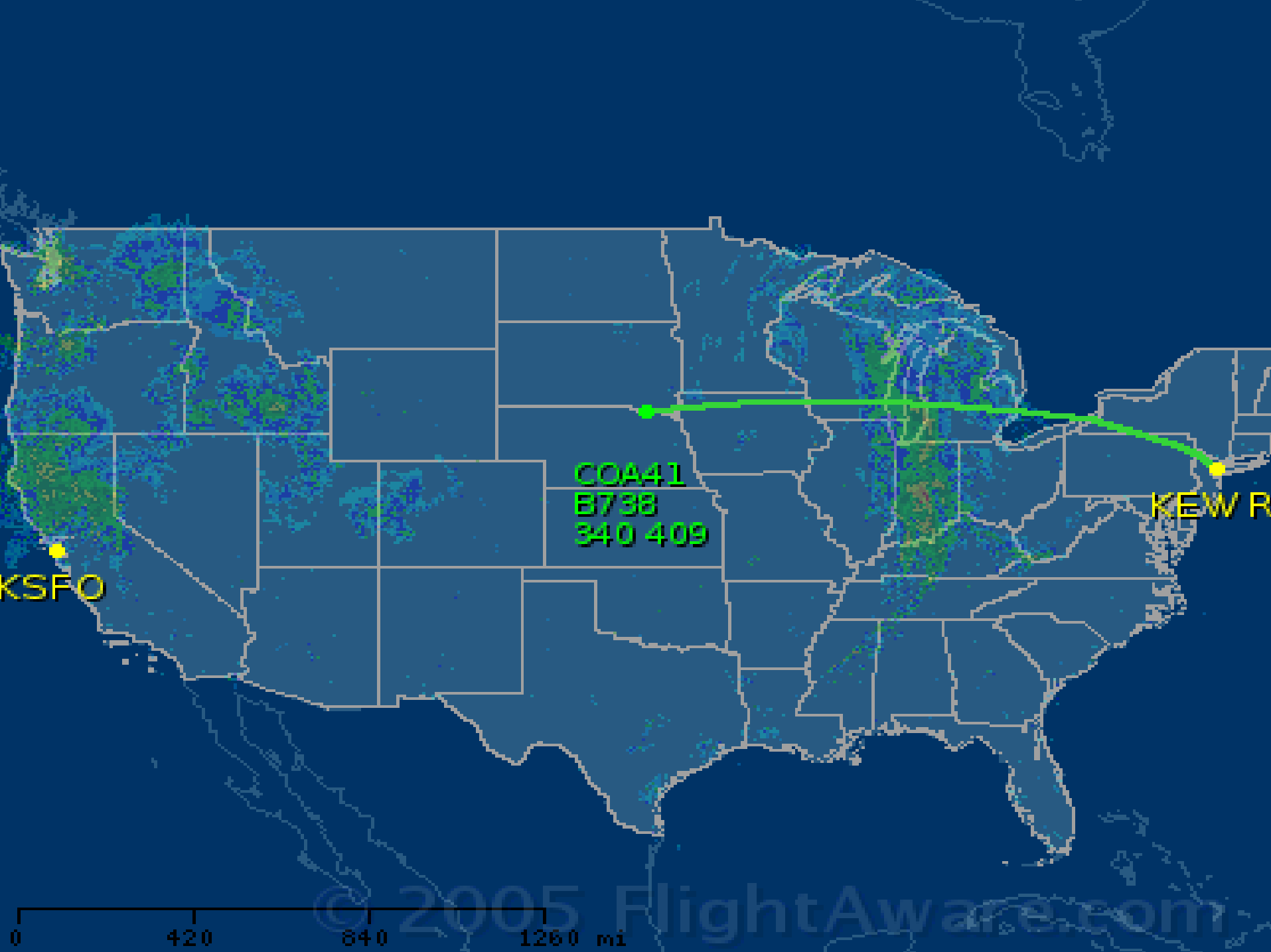
Date/Time	Location/Activity
Dec 19, 2005 9:28 AM	Ithaca, NY/Delivered
8:00 AM	ITHACA, NY/On FedEx vehicle for delivery
Dec 17, 2005 12:17 PM	ITHACA, NY/At local FedEx facility
9:26 AM	ITHACA, NY/At local FedEx facility
8:13 AM	SYRACUSE, NY/At dest sort facility
4:12 AM	MEMPHIS, TN/Departed FedEx location
12:03 AM	MEMPHIS, TN/Arrived at FedEx location
Dec 16, 2005 9:26 PM	WASHINGTON, DC/Left origin
6:58 PM	WASHINGTON, DC/Picked up
1:26 PM	/Package data transmitted to FedEx



5335 flights

Fri Jul 22 18:00:10 CDT 2005

© flightaware.com



KSFO

COA41
B738
340 409

KEWR

0 420 840 1260 mi

© 2005 FlightAware.com



West Summit of Snoqualmie Pass WSDOT ©

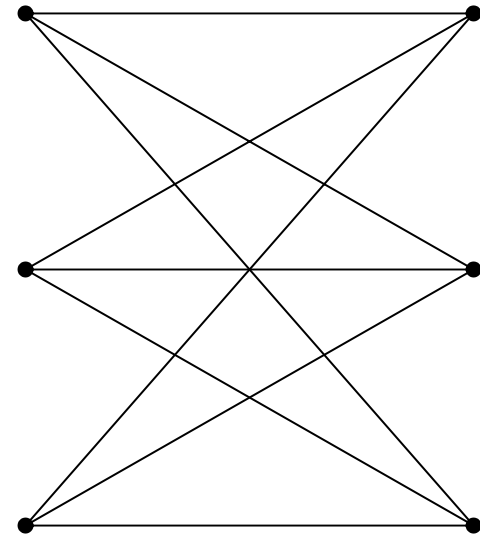
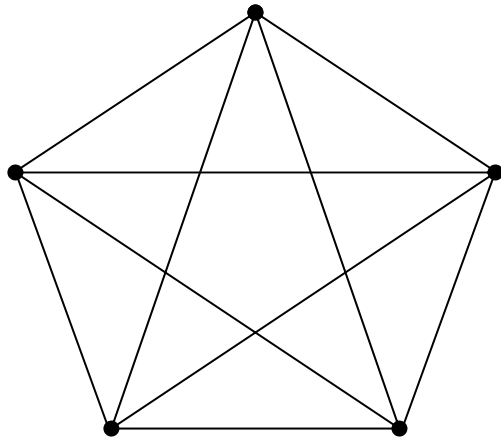




Theory to support new directions

- Large graphs
- Spectral analysis
- High dimensions and dimension reduction
- Clustering
- Collaborative filtering
- Extracting signal from noise

Graph Theory of the 50's



Theorem: A graph is planar if it does not contain a Kuratowski subgraph as a contraction.

Theory of Large Graphs

Large graphs

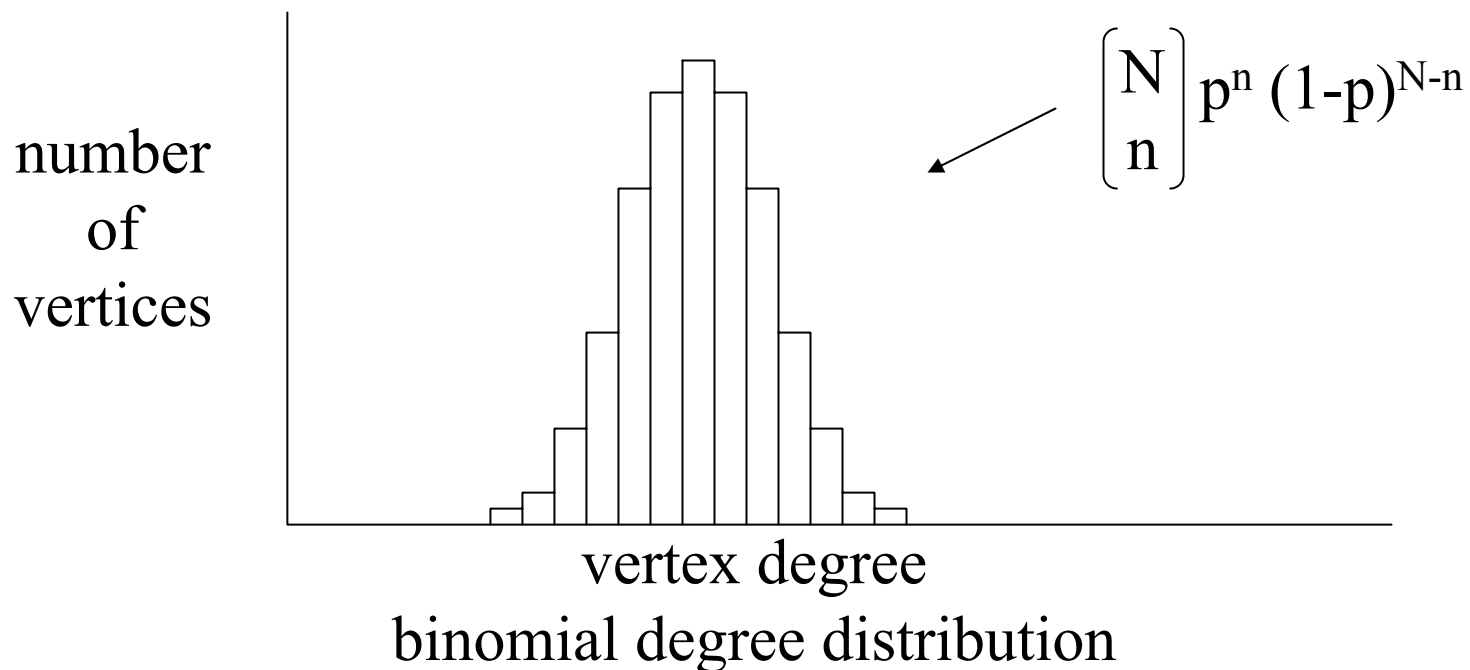
- Billion vertices
- Exact edges present not critical

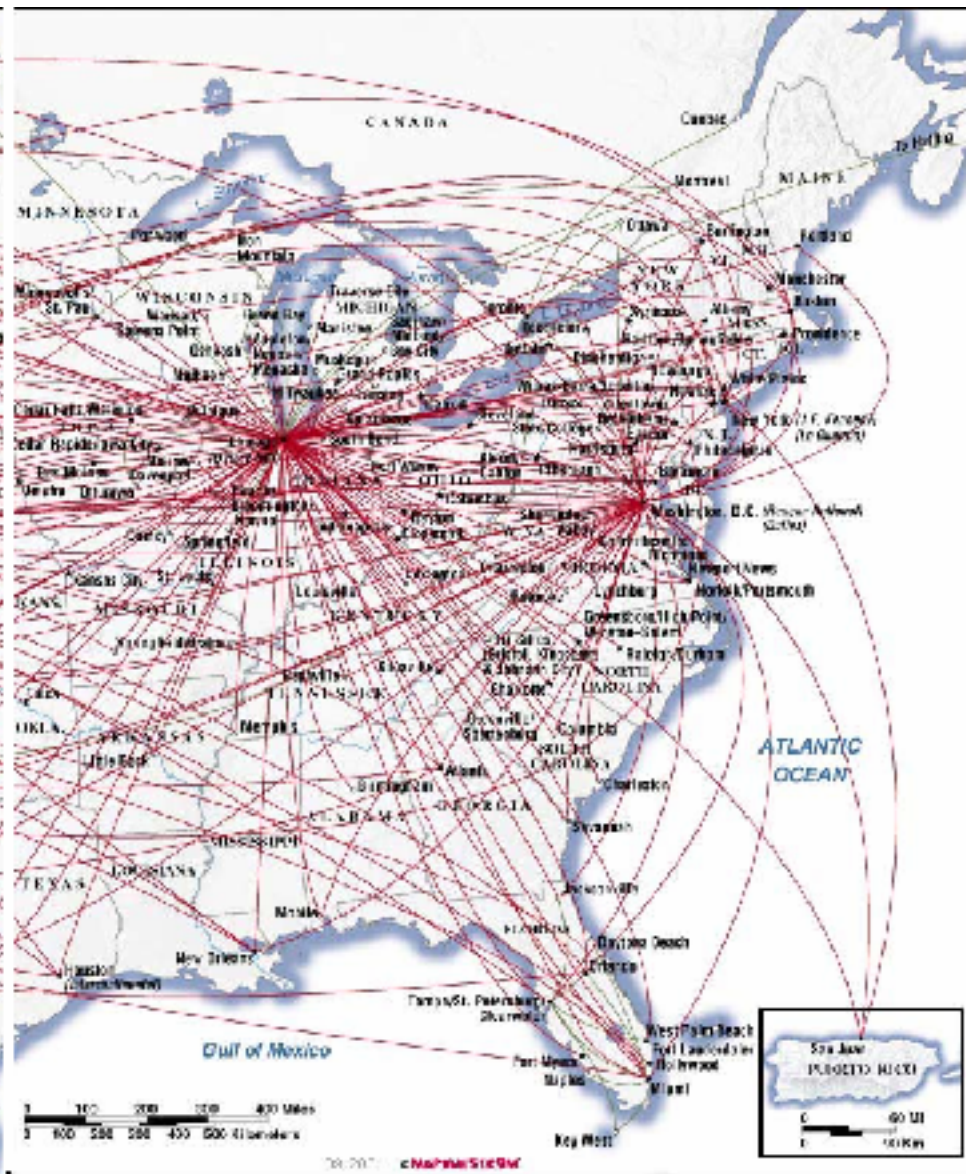
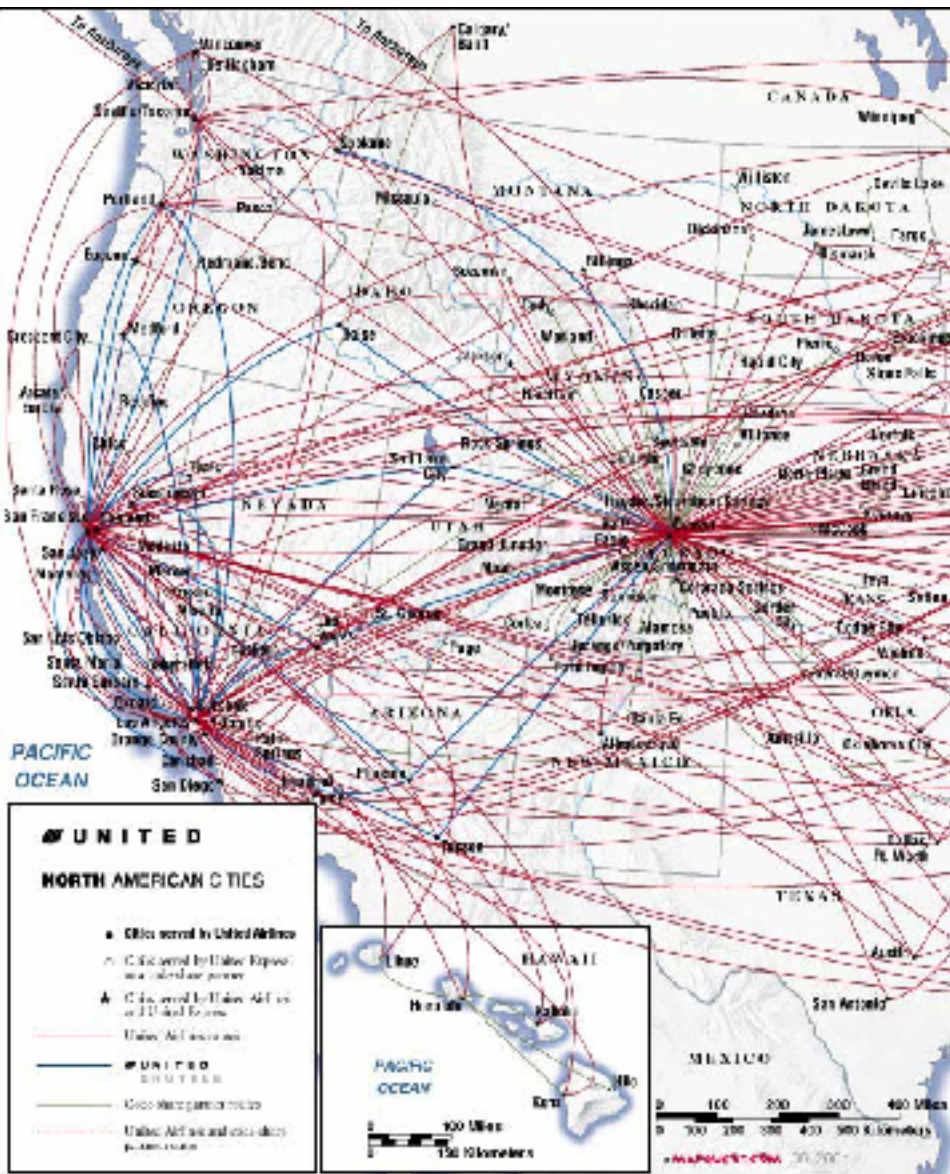
Theoretical basis for study of large graphs

- Maybe theory of graph generation
- Invariant to small changes in definition
- Must be able to prove basic theorems

Erdős-Renyi

- n vertices
- each of n^2 potential edges is present with independent probability





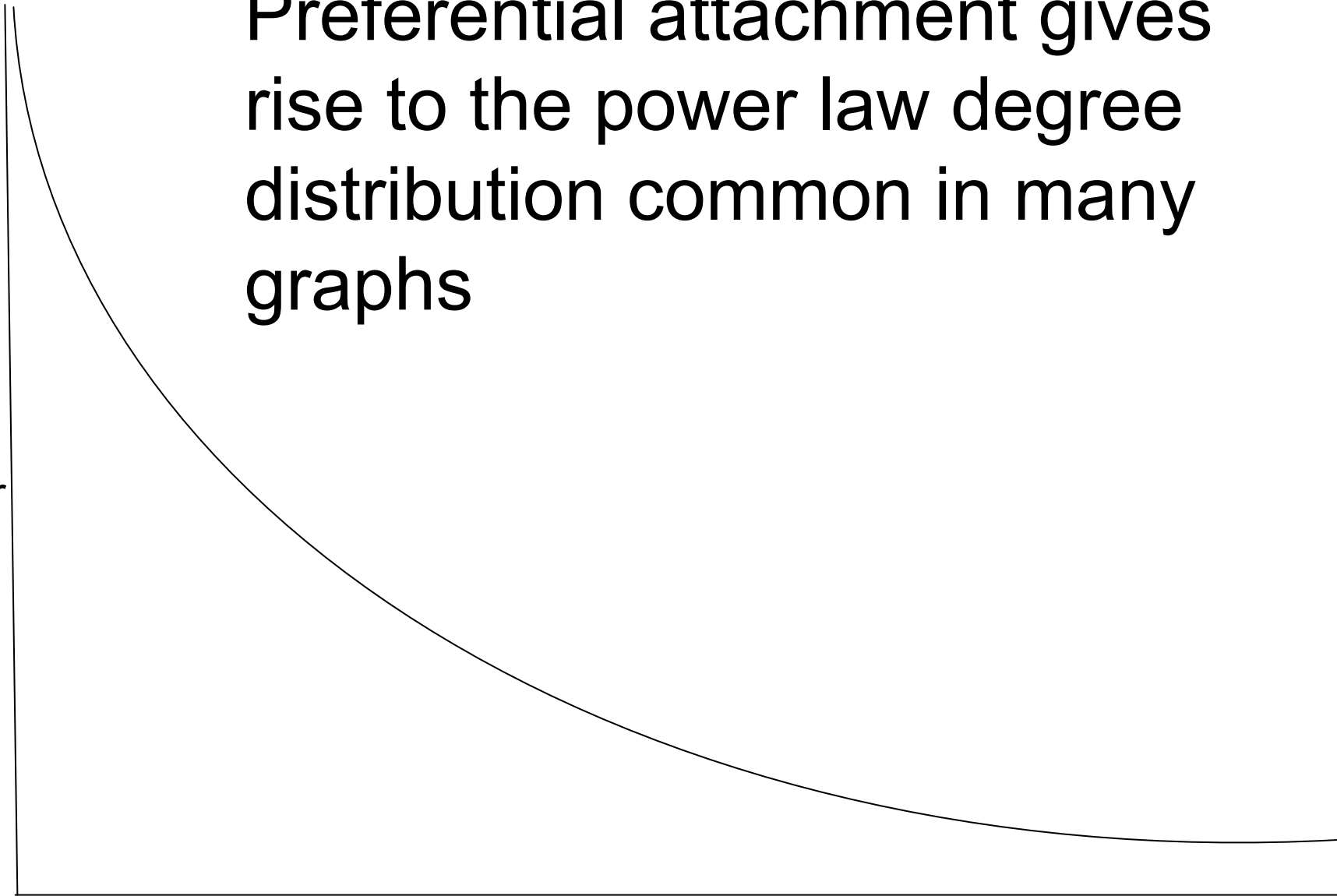
Generative models for graphs

- Vertices and edges added at each unit of time
- Rule to determine where to place edges
 - Uniform probability
 - Preferential attachment - gives rise to power law degree distributions

Preferential attachment gives rise to the power law degree distribution common in many graphs

Number
of
vertices

Vertex degree



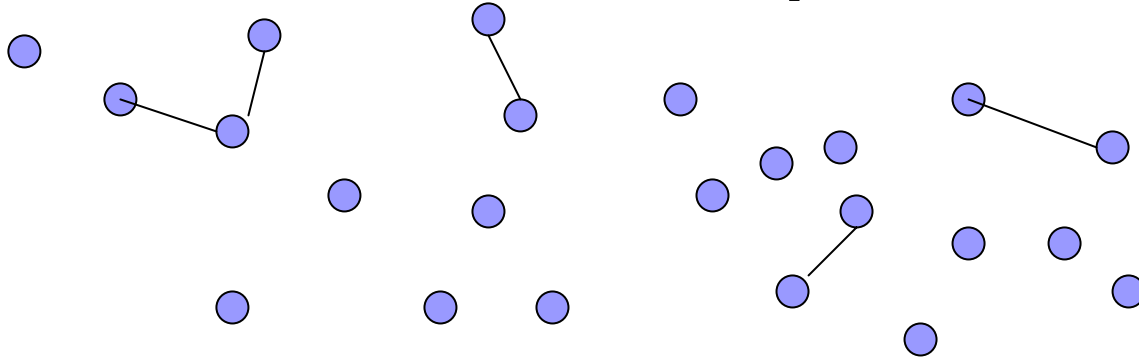
Protein interactions

2730 proteins in data base

3602 interactions between proteins

SIZE OF COMPONENT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...	1851
NUMBER OF COMPONENTS	48	179	50	25	14	6	4	6	1	1	1	0	0	0	0	1		1

Giant Component



1. Create n isolated vertices
2. Add Edges randomly one by one
3. Compute number of connected components

Giant Component

1	
1000	
1	2
998	1

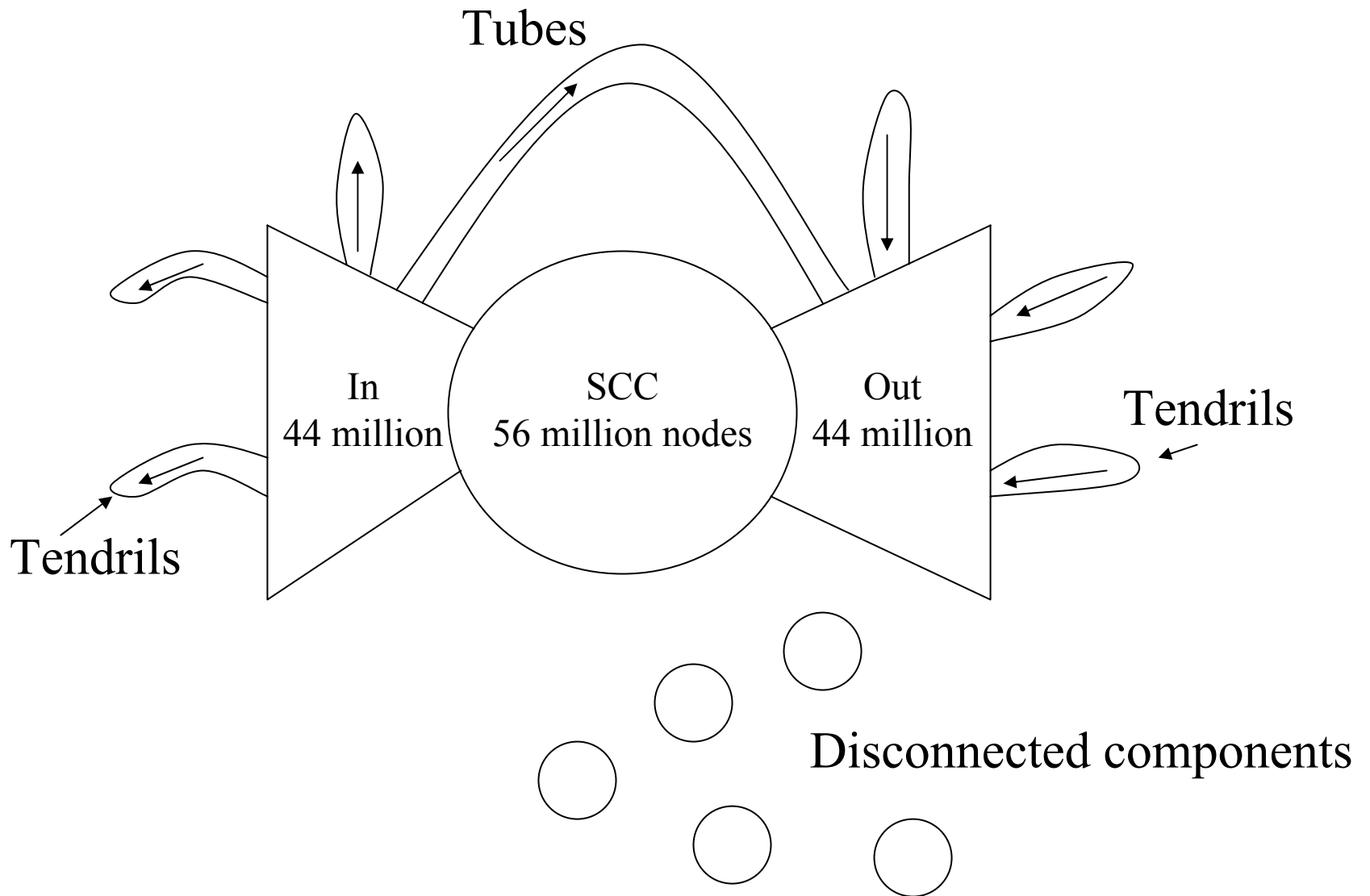
1	2	3	4	5	6	7	8	9	10	11
548	89	28	14	9	5	3	1	1	1	1

Giant Component

1	2	3	4	5	6	7	8	9	10	11
548	89	28	14	9	5	3	1	1	1	1

1	2	3	4	5	6	7	8
367	70	24	12	9	3	2	2
9	10	12	13	14	20	55	101
2	2	1	2	2	1	1	1

1	2	3	4	5	6	7	8	9	11	514
252	39	13	6	3	6	2	1	1	1	1



Phase transitions

■ $G(n,p)$

- Emergence of cycle
- Giant component
- Connected graph

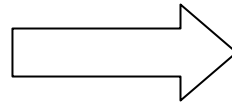
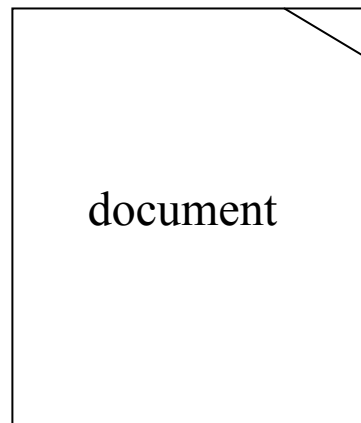
■ $N(p)$

- Emergence of arithmetic sequence

■ CNF satisfiability

- Fix number of variables, increase number of clauses

Access to Information SMART Technology



aardvark	0
abacus	0
⋮	
antitrust	42
⋮	
CEO	17
⋮	
microsoft	61
⋮	
windows	14
wine	0
wing	0
winner	3
winter	0
⋮	
zoo	0
zoology	0
Zurich	0

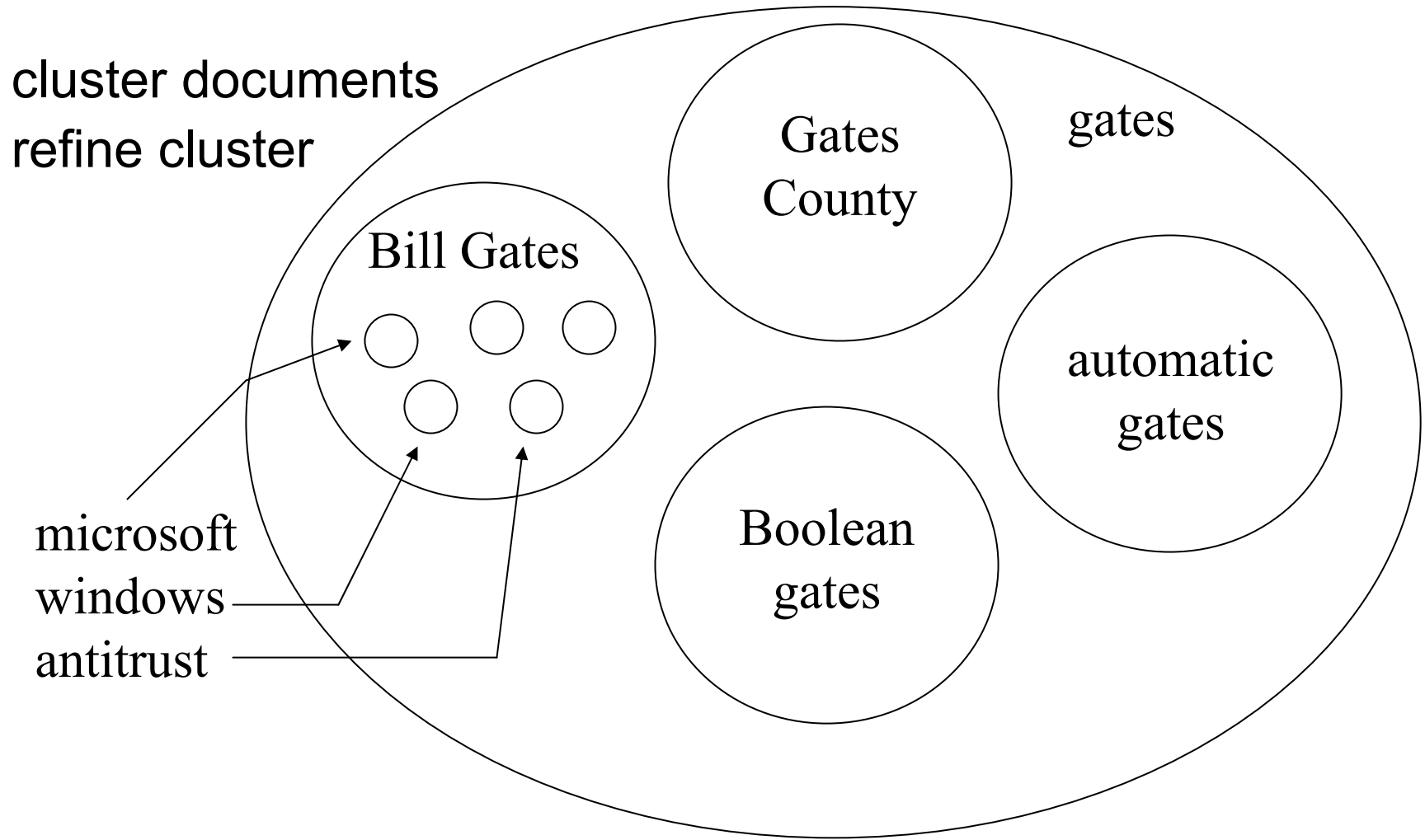
Locating relevant documents

Query: Where can I get information on gates?

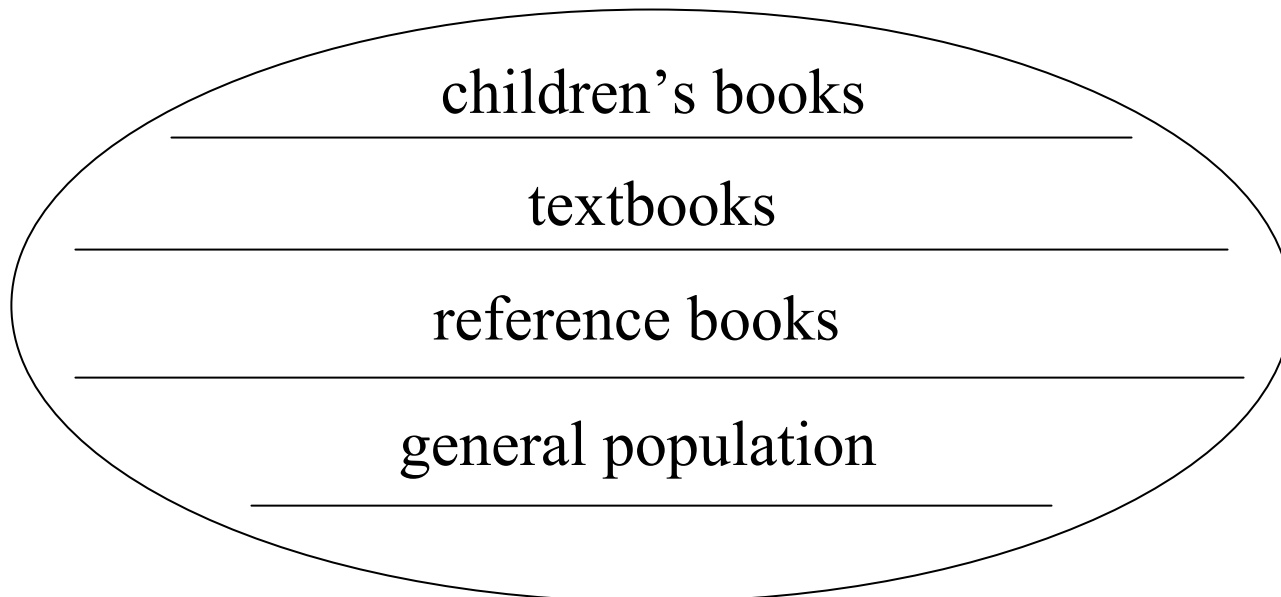
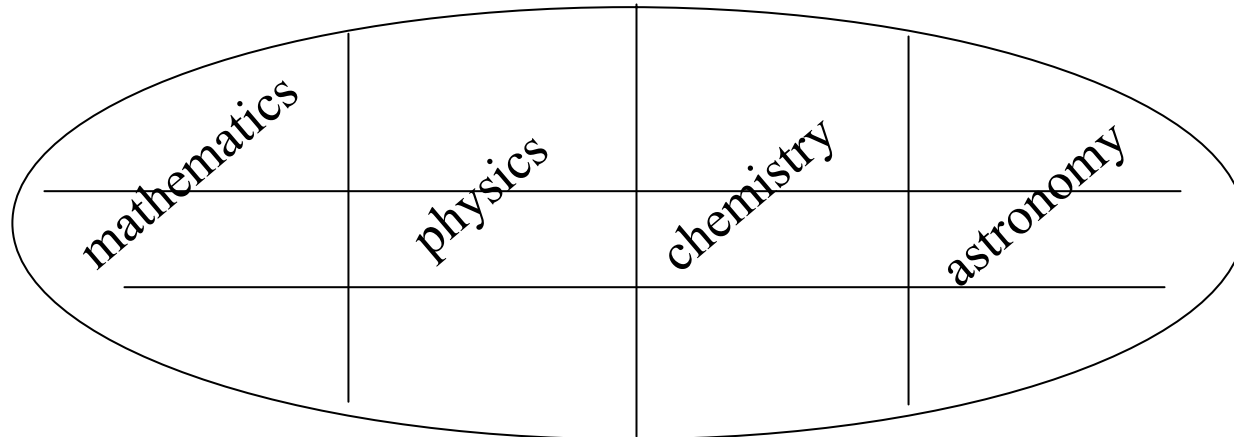
2,060,000 hits

Bill Gates	593,000
Gates county	177,000
baby gates	170,000
gates of heaven	169,000
automatic gates	83,000
fences and gates	43,000
Boolean gates	19,000

Clustering documents

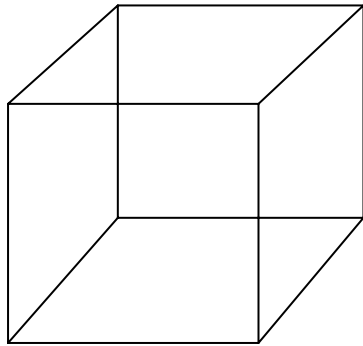


Refinement of another type: books

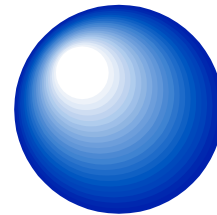


High Dimensions

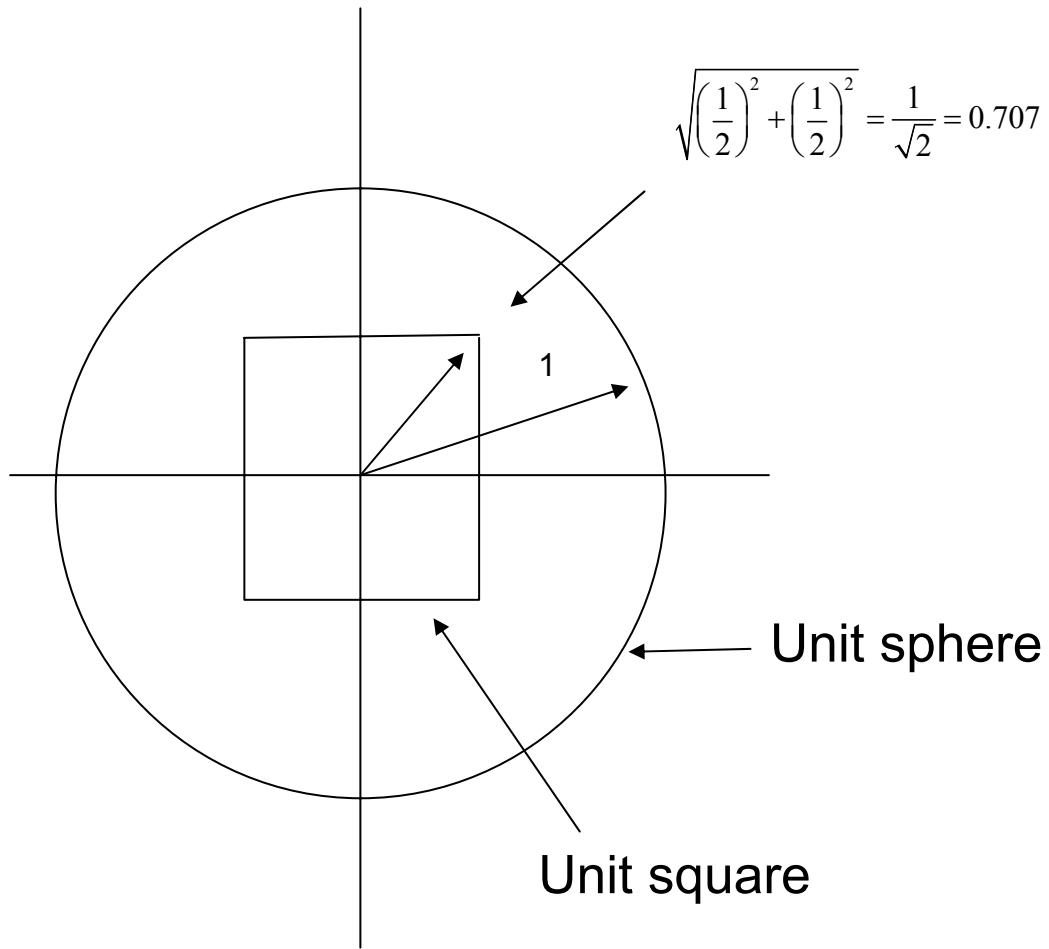
Intuition from two and three dimensions not valid for high dimension



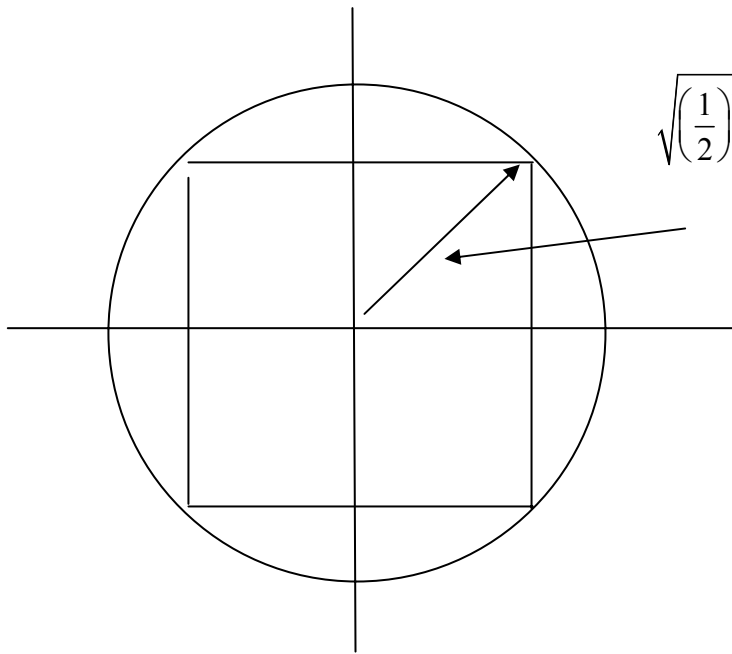
Volume of cube is
one in all
dimensions



Volume of
sphere goes to
zero

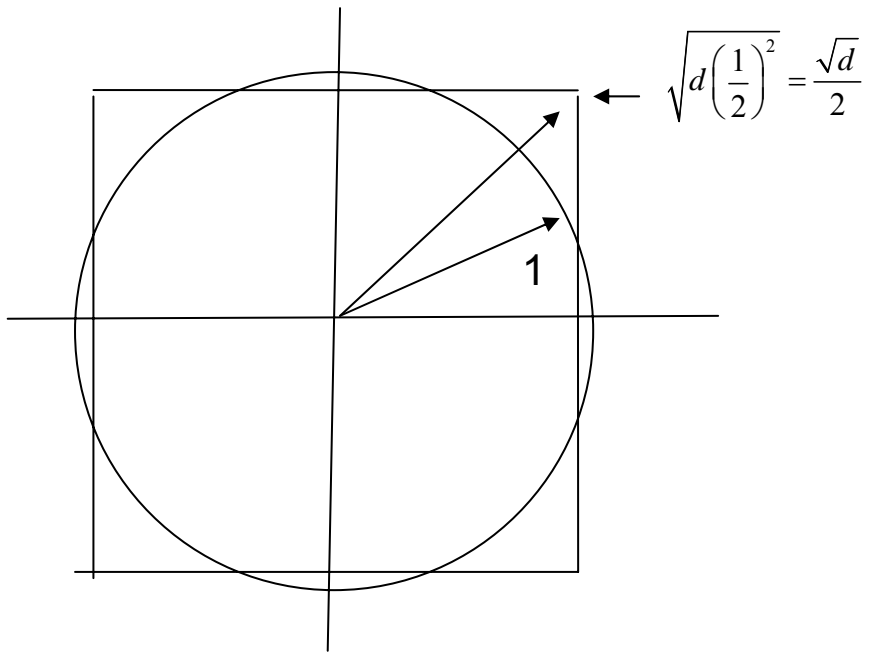


2 Dimensions



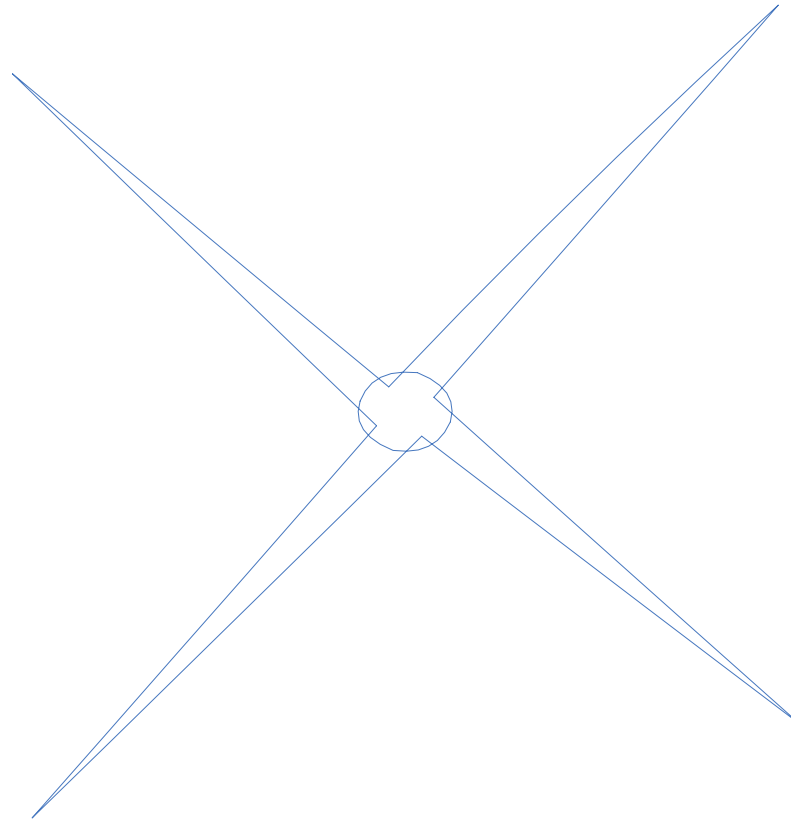
$$\sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2} = 1$$

4 Dimensions

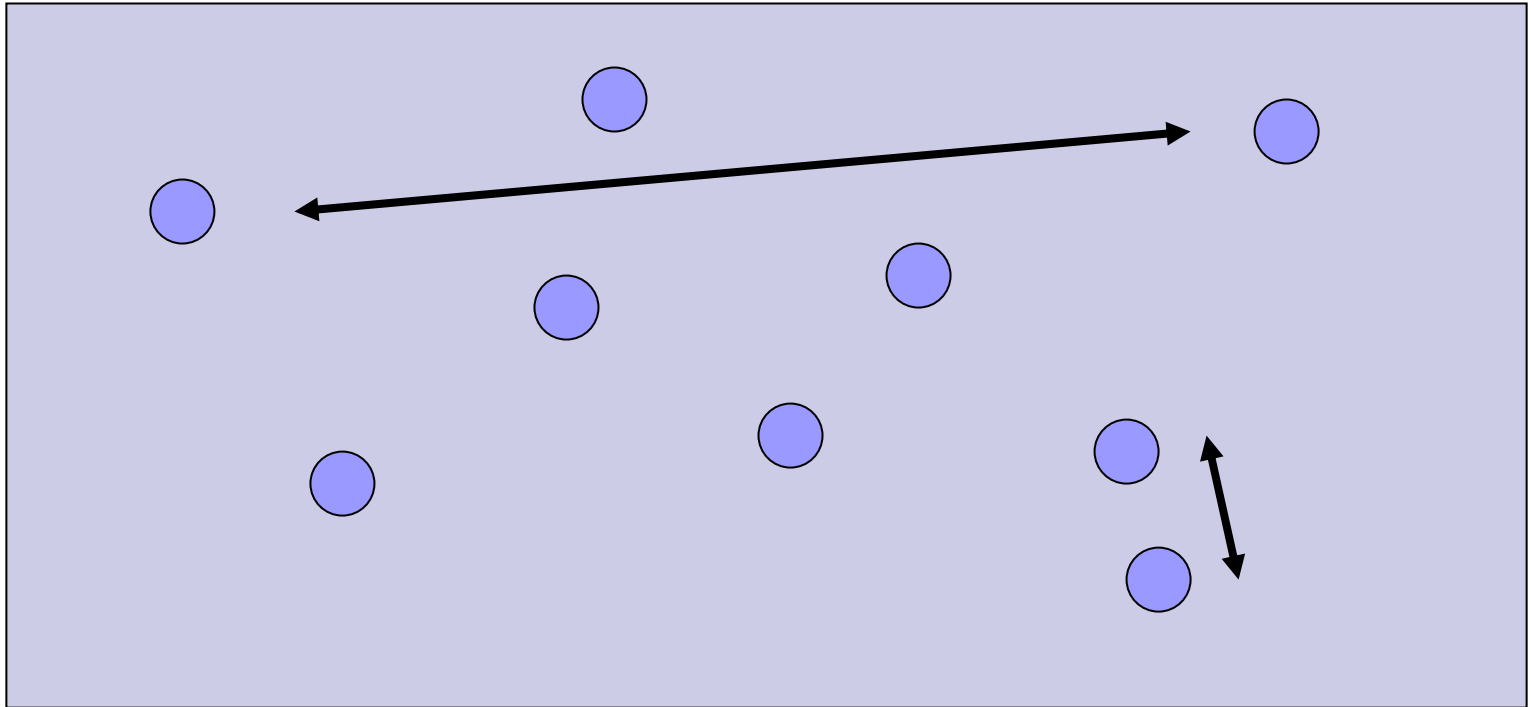


d Dimensions

Almost all area of the unit cube is
outside the unit sphere



High dimension is fundamentally different from 2 or 3 dimensional space

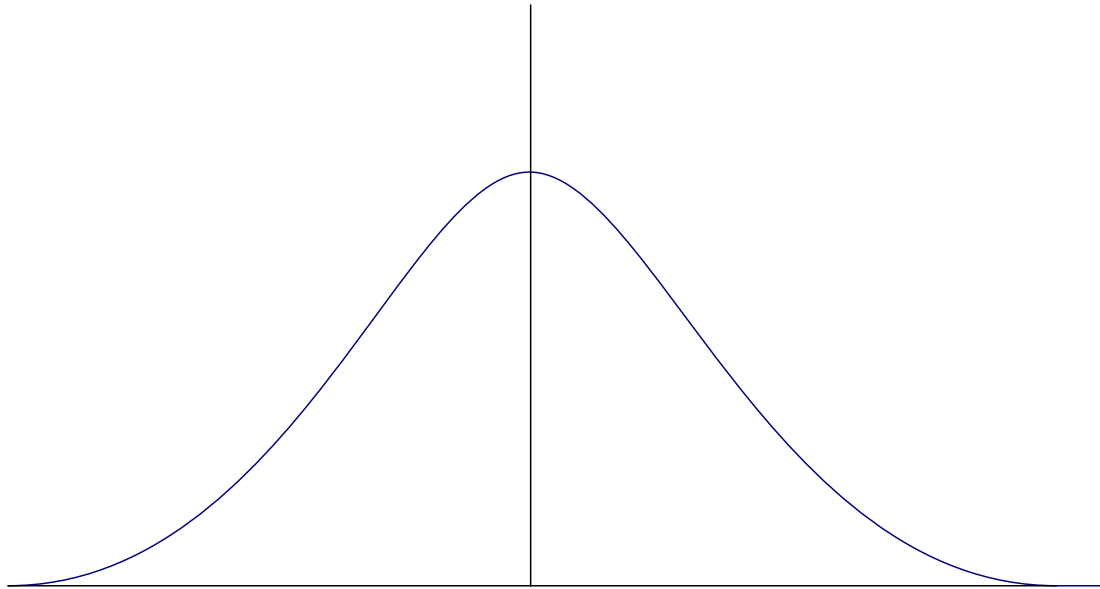


High dimensional data is inherently unstable

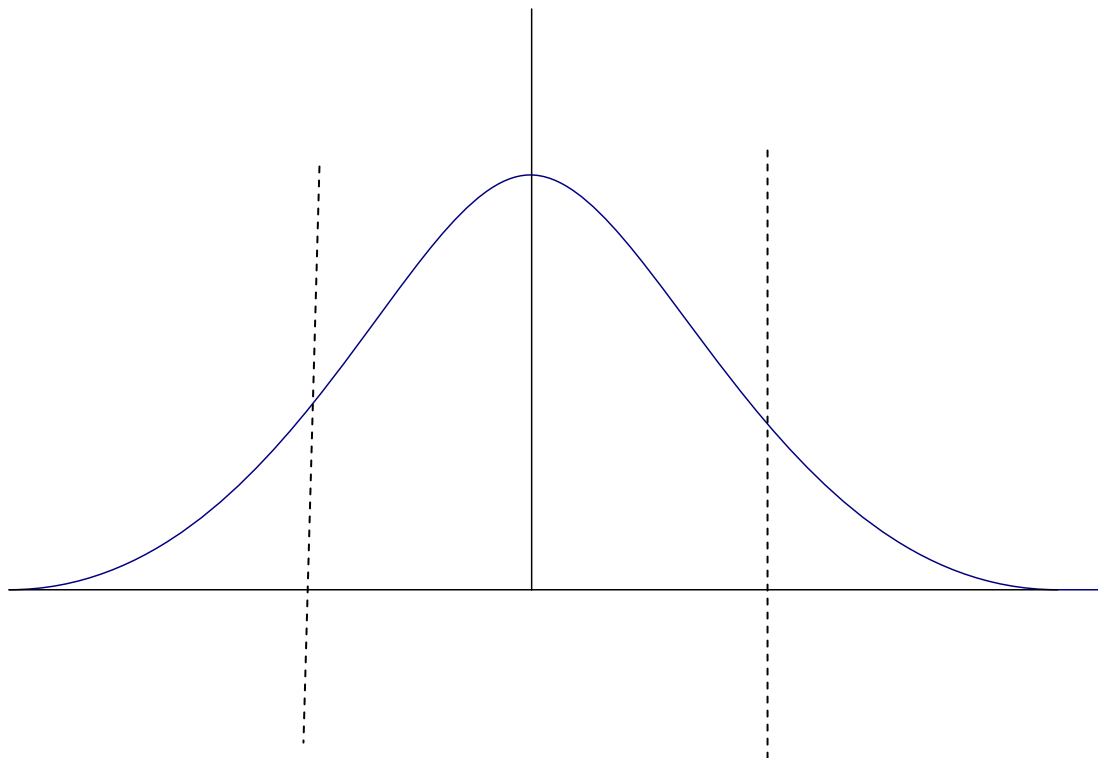
- Given n random points in d dimensional space essentially all n^2 distances are equal.

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

Gaussian distribution

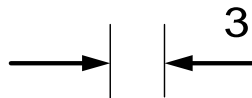
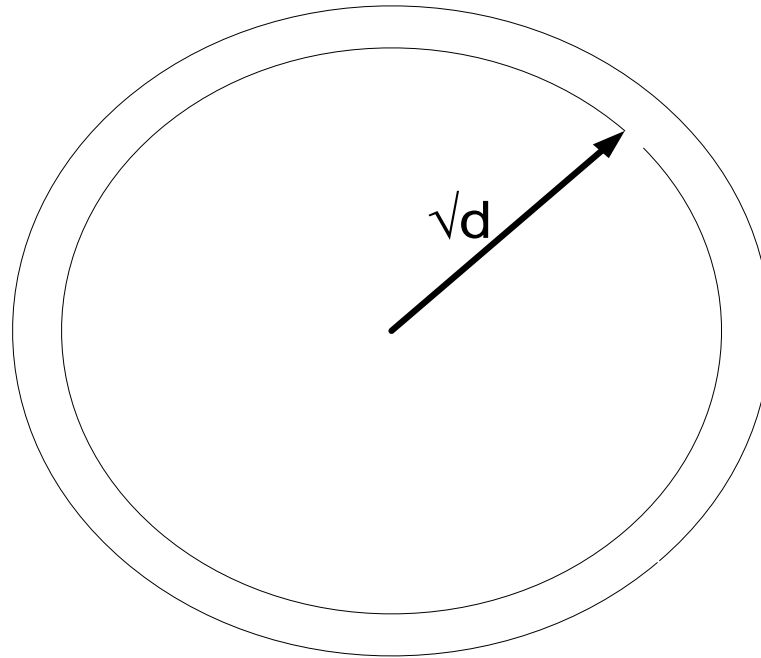


Gaussian distribution

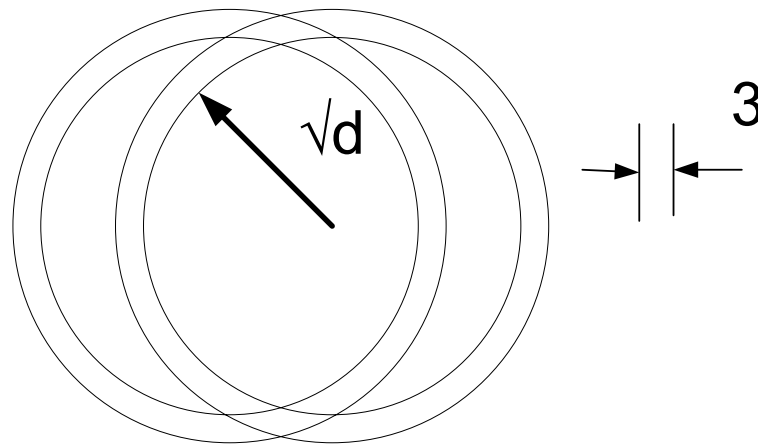


Probability mass concentrated
between dotted lines

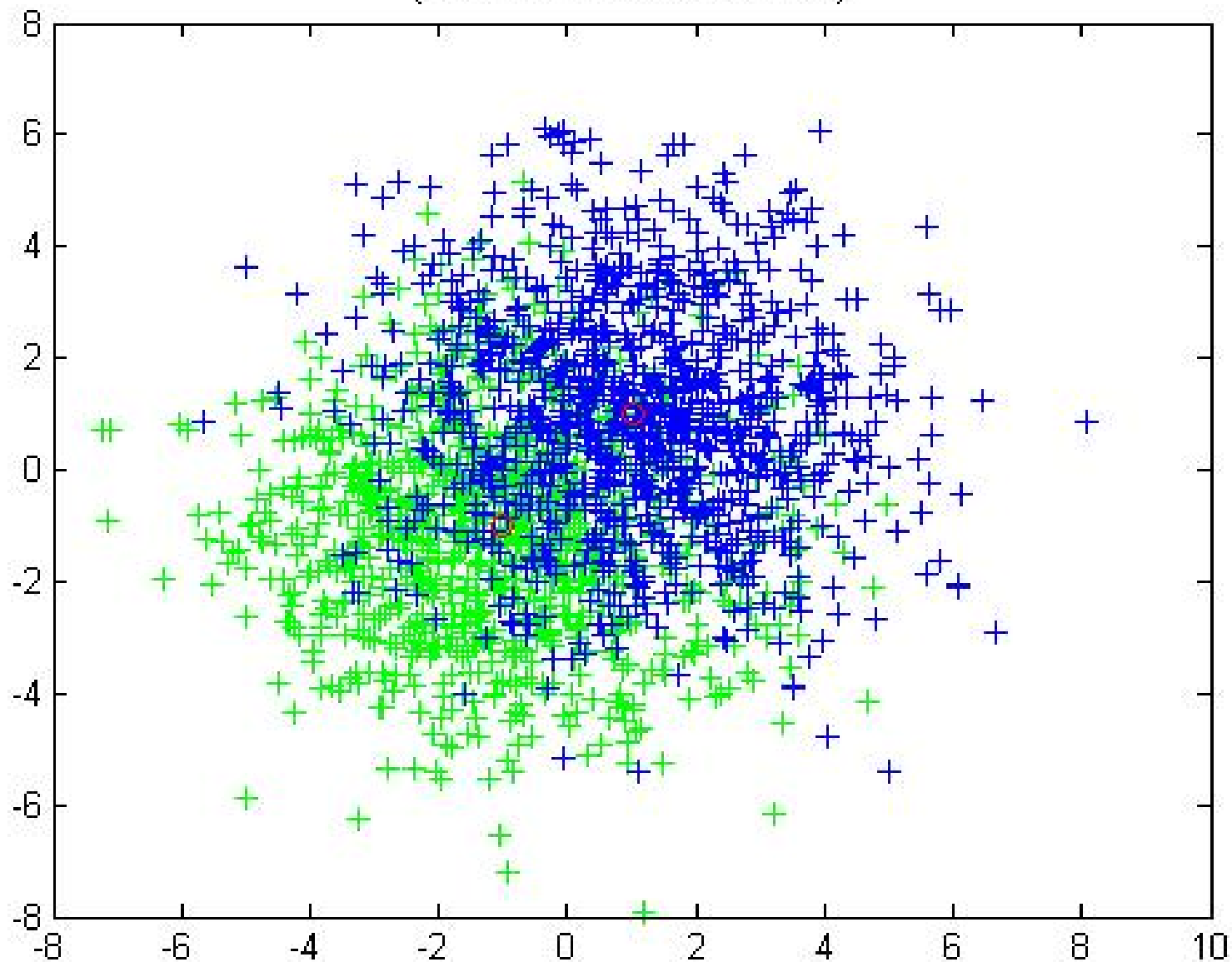
Gaussian in high dimensions



Two Gaussians

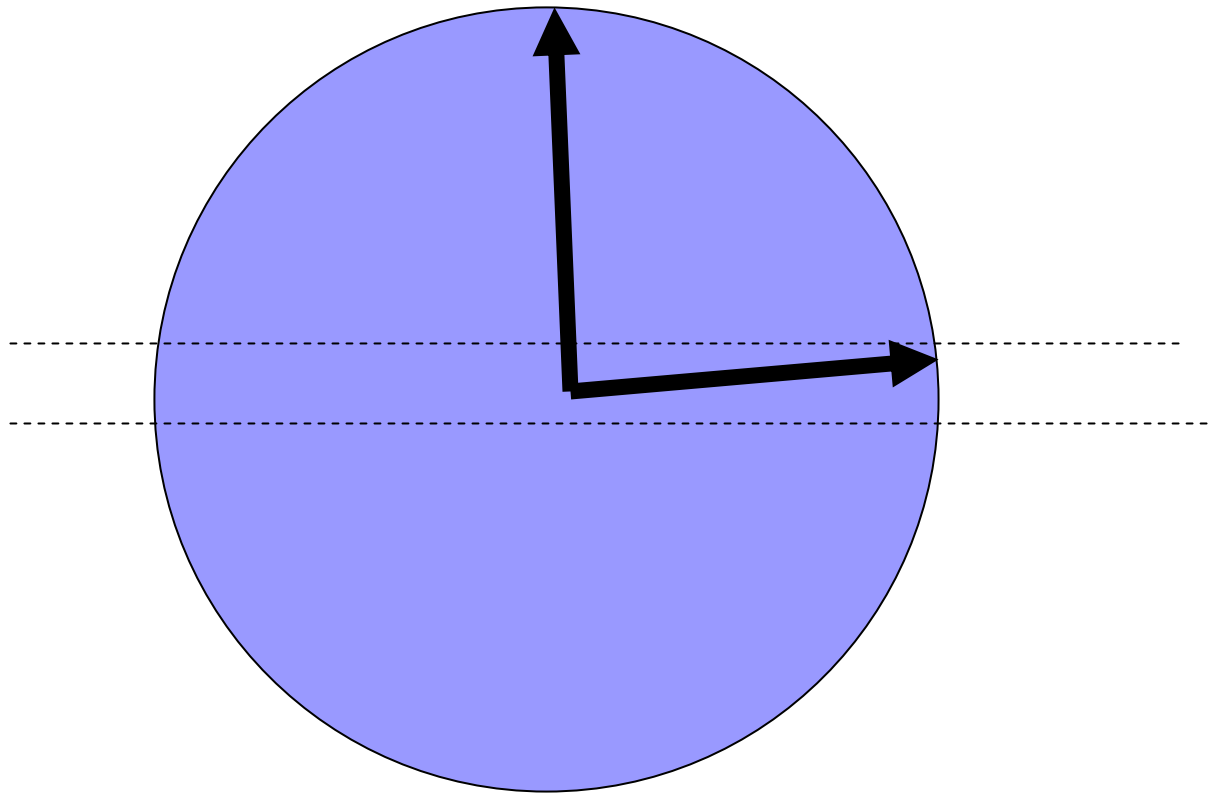


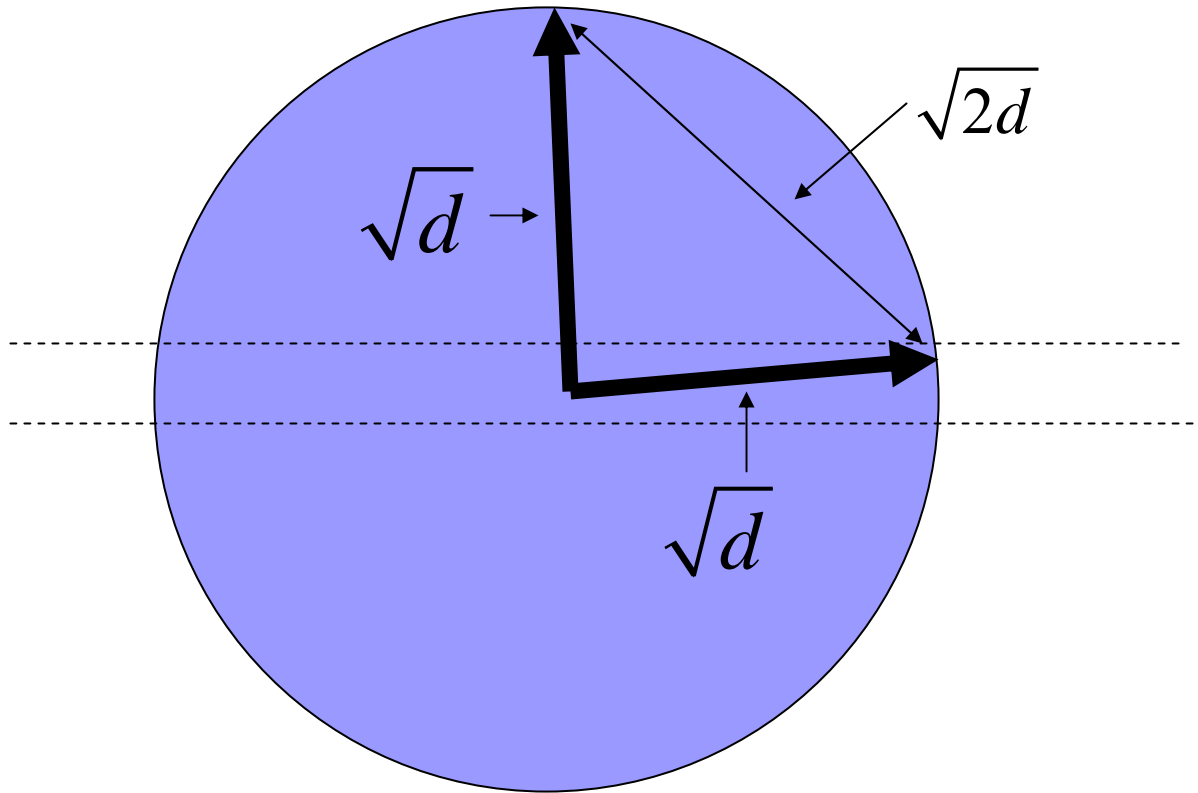
2 Gaussians with 1000 points each: $\mu=1.000000$, $\sigma=2.000000$, $\text{dim}=500$
(Dimensions 1 and 2 shown)



Distance between two random points from same Gaussian

- Points on thin annulus of radius \sqrt{d}
- Approximate by sphere of radius \sqrt{d}
- Average distance between two points is \sqrt{d}
(Place one pt at N. Pole other at random.
Almost surely second point near the equator.)





Can separate points from two Gaussians if

$$\sqrt{\delta^2 + 2d} > \sqrt{2d} + \gamma$$

$$\sqrt{2d} \left(1 + \frac{1}{2} \frac{\delta^2}{2d} + \dots\right) > \sqrt{2d} + \gamma$$

$$\frac{1}{2} \frac{\delta^2}{\sqrt{2d}} > \gamma$$

$$\delta > \sqrt{2\gamma} (2d)^{\frac{1}{4}}$$

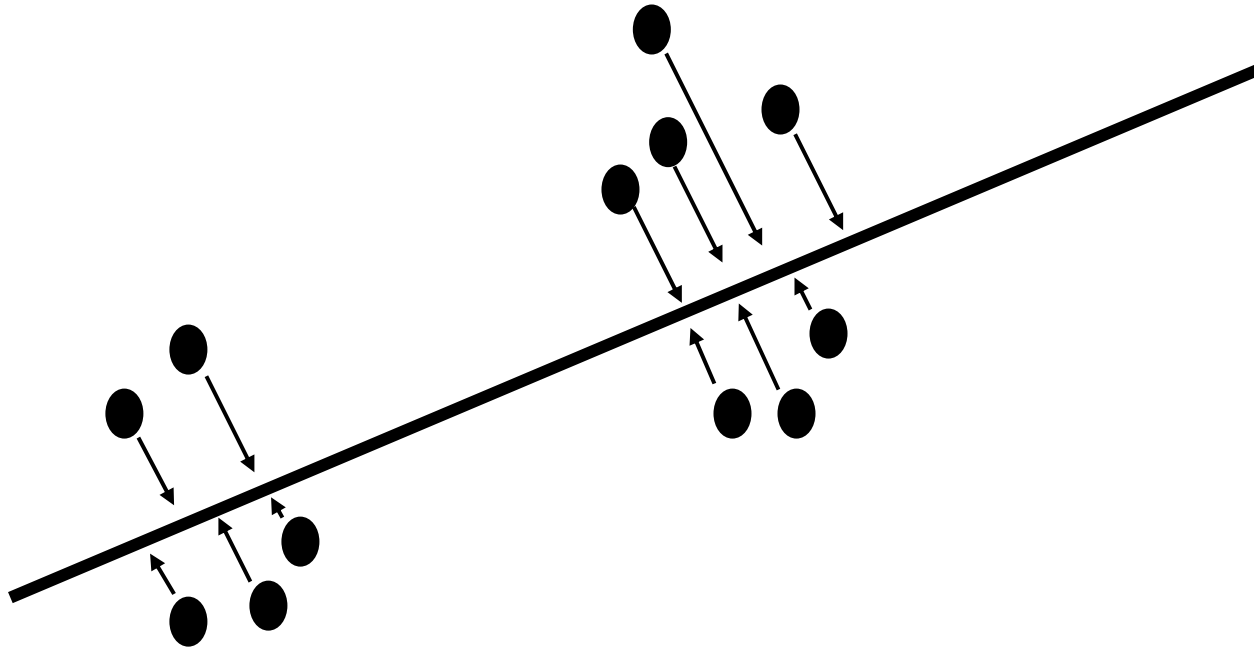


Dimension reduction

- Project points onto subspace containing centers of Gaussians
- Reduce dimension from d to k , the number of Gaussians

- Centers retain separation


- Average distance between points reduced by $\sqrt{\frac{d}{k}}$

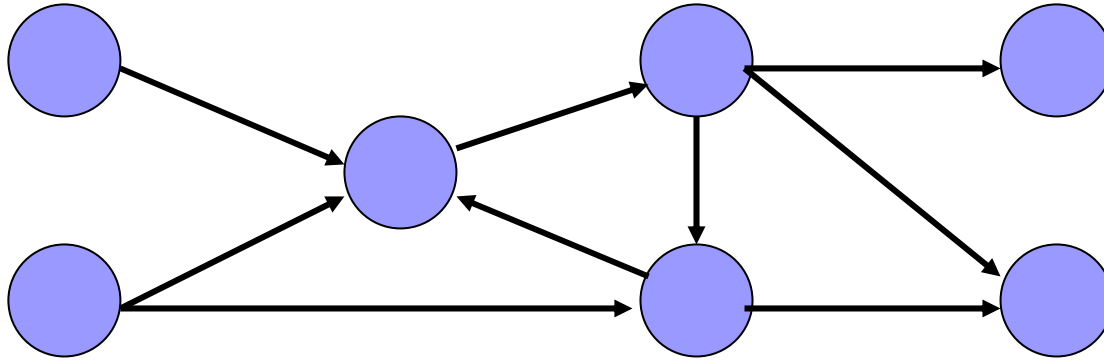


Can separate Gaussians provided

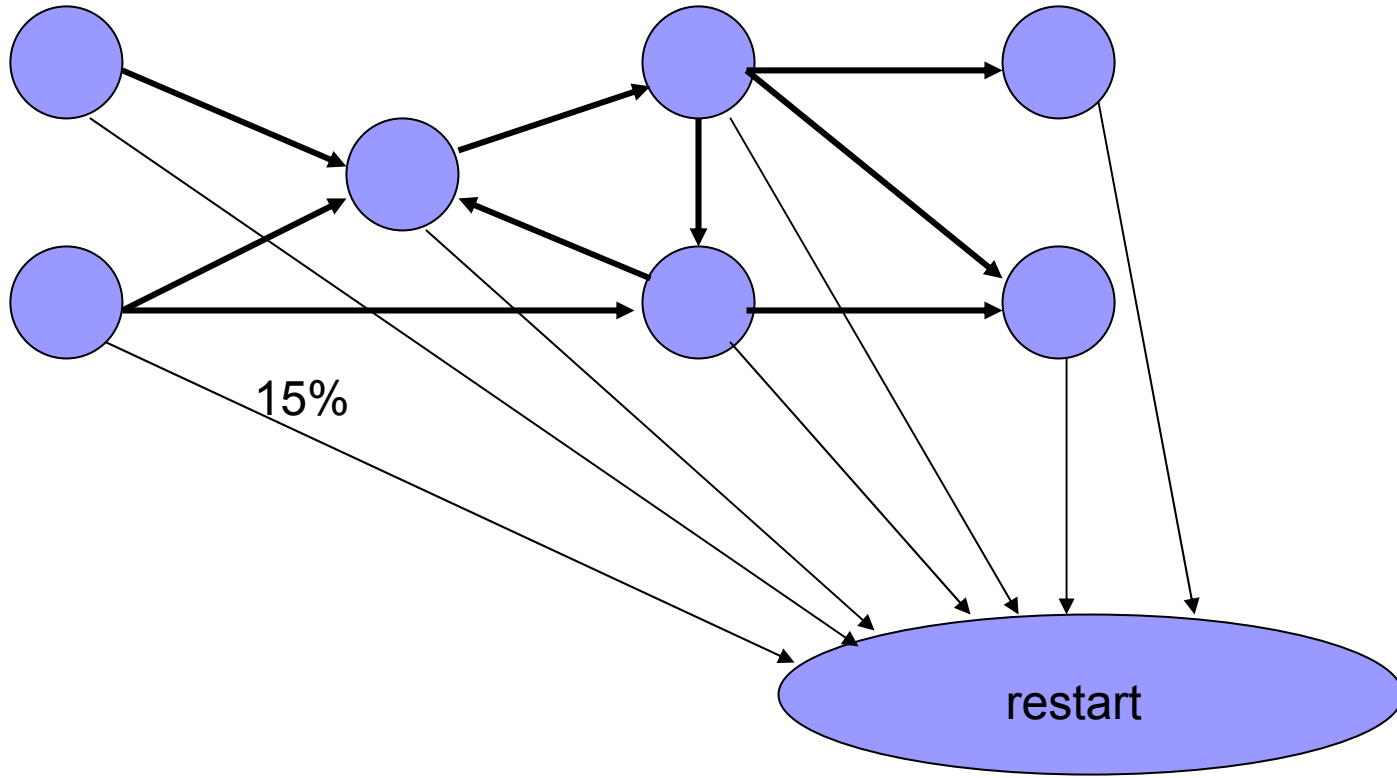
$$\sqrt{\delta^2 + 2k} > \sqrt{2k} + \gamma$$

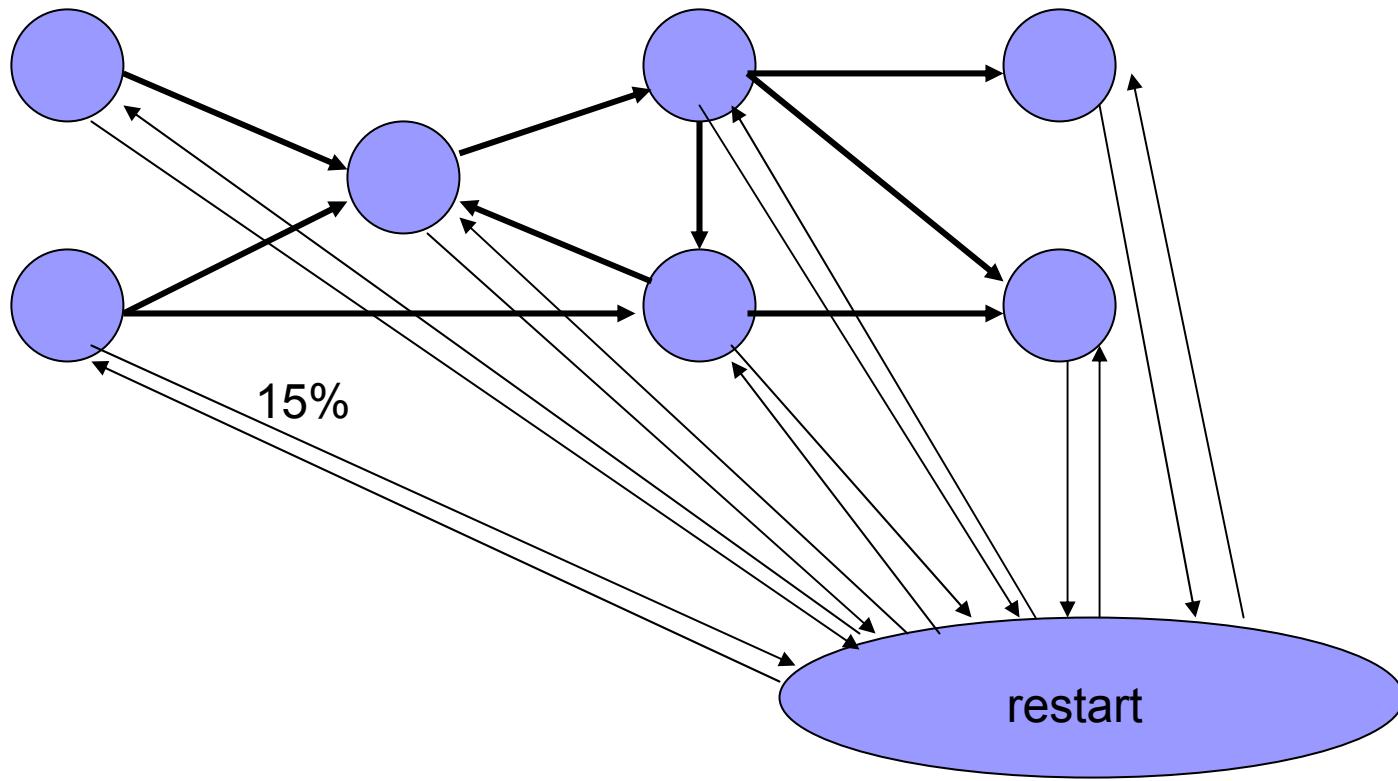
δ > some constant involving k and γ
independent of the dimension

- 
- Ranking is important
 - Restaurants
 - Movies
 - Web pages
 - Multi billion dollar industry



Page rank equals stationary probability of random walk





Restart yields strongly connected graph

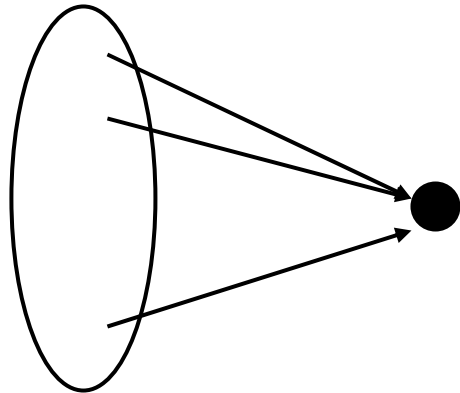


Suppose you wish to increase the page rank of vertex v

- Capture restart
web farm
- Capture random walk
small cycles

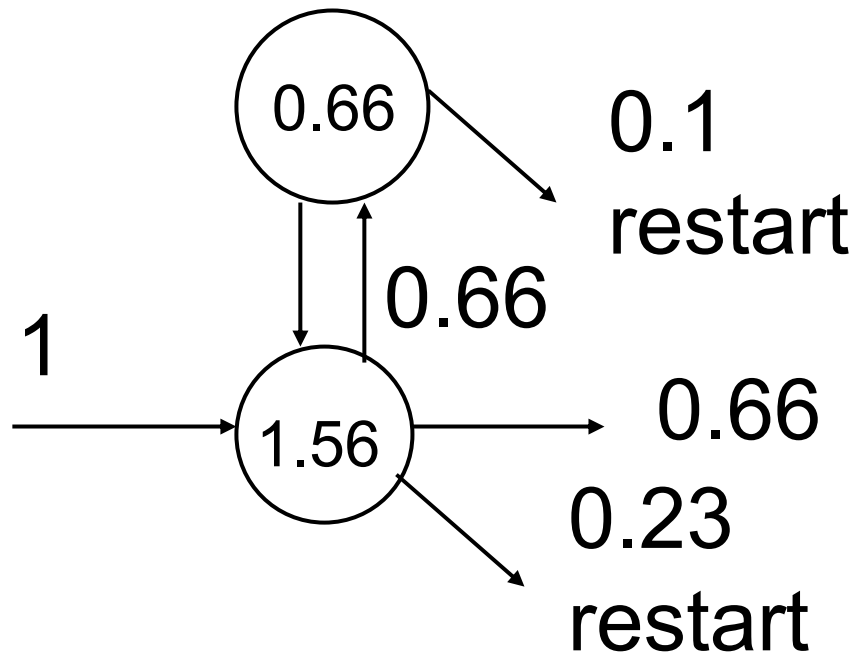
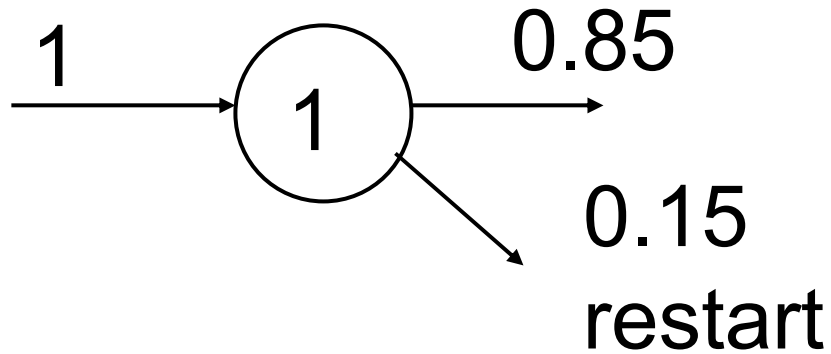
Capture restart

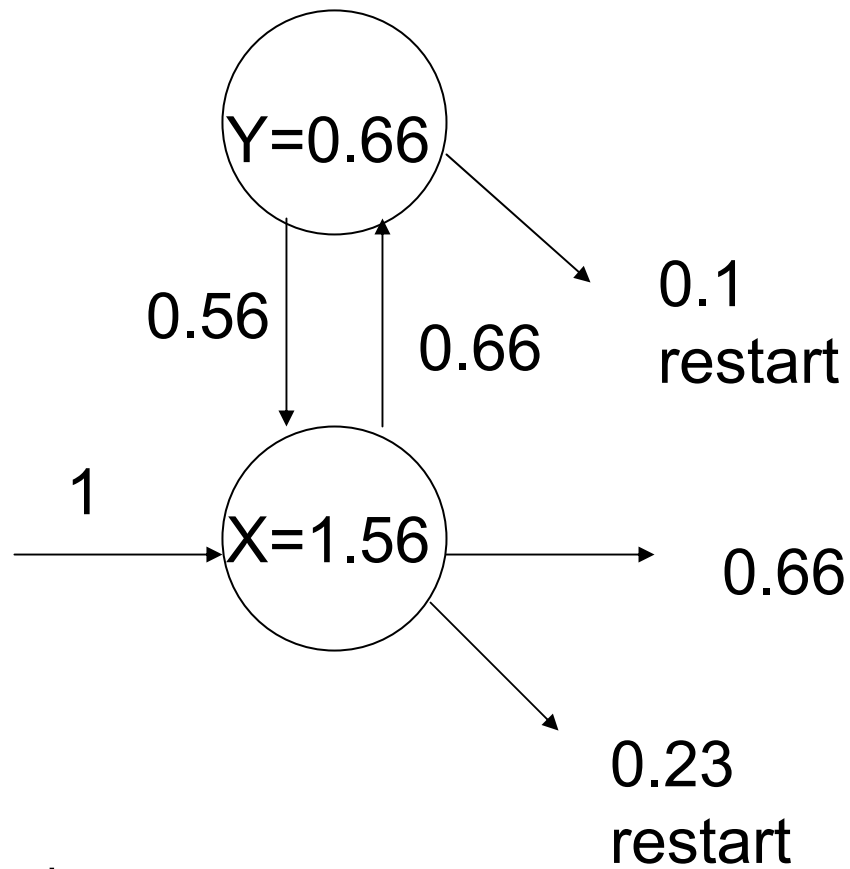
- Buy 20,000 url's and capture restart



- Can be countered by small restart value
- Small restart increases web rank of page that captures random walk by small cycles.


Capture random walk





$$X=1+0.85*y$$

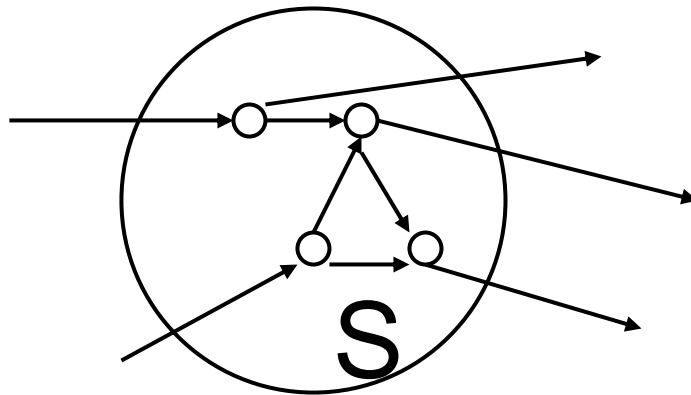
$$Y=0.85*x/2$$



If one loop increases Pagerank from 1 to 1.56 why not add many self loops?

Maximum increase in Pagerank is 6.67

Discovery time – time to first reach a vertex by random walk from uniform start from uniform start



Cannot lower discovery time of any page in S below minimum already in S




Why not replace Pagerank by
discovery time?

No efficient algorithm for
discovery time

DiscoveryTime(v)


remove edges out of v


calculate Pagerank(v) in
modified graph

- 
- Is there a way for a spammer to raise Pagerank in a way that is not statistically detectable

Information is important

- When a customer makes a purchase what else is he likely to buy?
 - Camera
 - Memory card
 - Batteries
 - Carrying case
 - Etc.
- Knowing what a customer is likely to buy is important information.

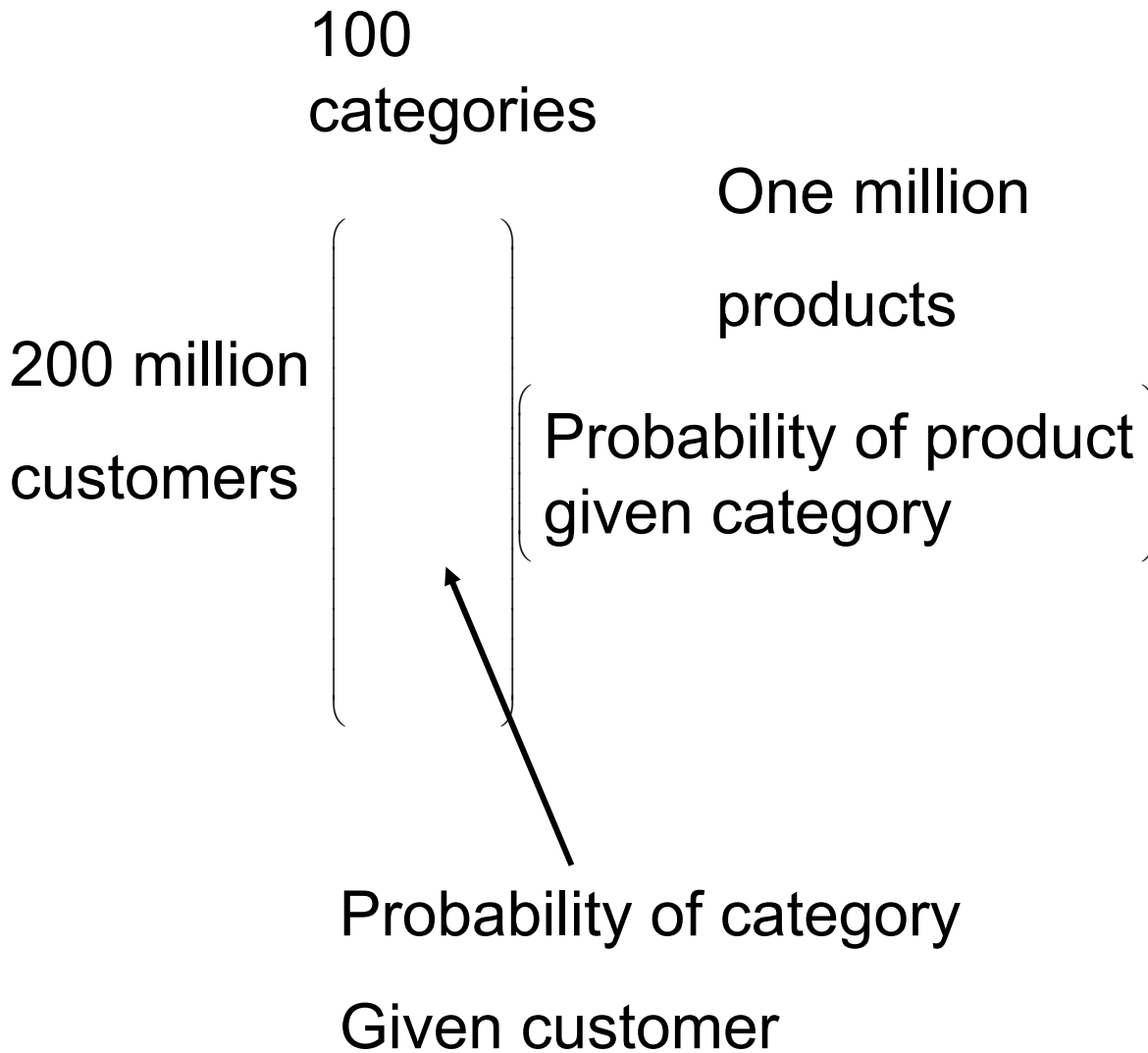
- 
- How can we extract information from a customer's visit to a web site?
 - What web pages were visited?
 - What order?
 - How long?

- 
- Collaborative filtering
 - Recommendations
 - Which pop-up ads
 - Detecting changes over time
 - Changes in a market
 - Buying habits
 - Access to information

One million products

200 million
customers

Probability of customer
Buying product





Extracting Information from Large Data Sources

- Data streams
- Large data collections
- Detecting changes in patterns



Detecting trends before they become obvious

- Is some category of customer changing their buying habits?
 - Purchases, travel destination, vacations
- Is there some new trend in the stock market?
- How do we detect changes in a large database over time?



Identifying Changing Patterns in a Large Data Set

- How soon can one detect a change in patterns in a large volume of information?
- How large must a change be in order to distinguish it from random fluctuations?



Conclusions

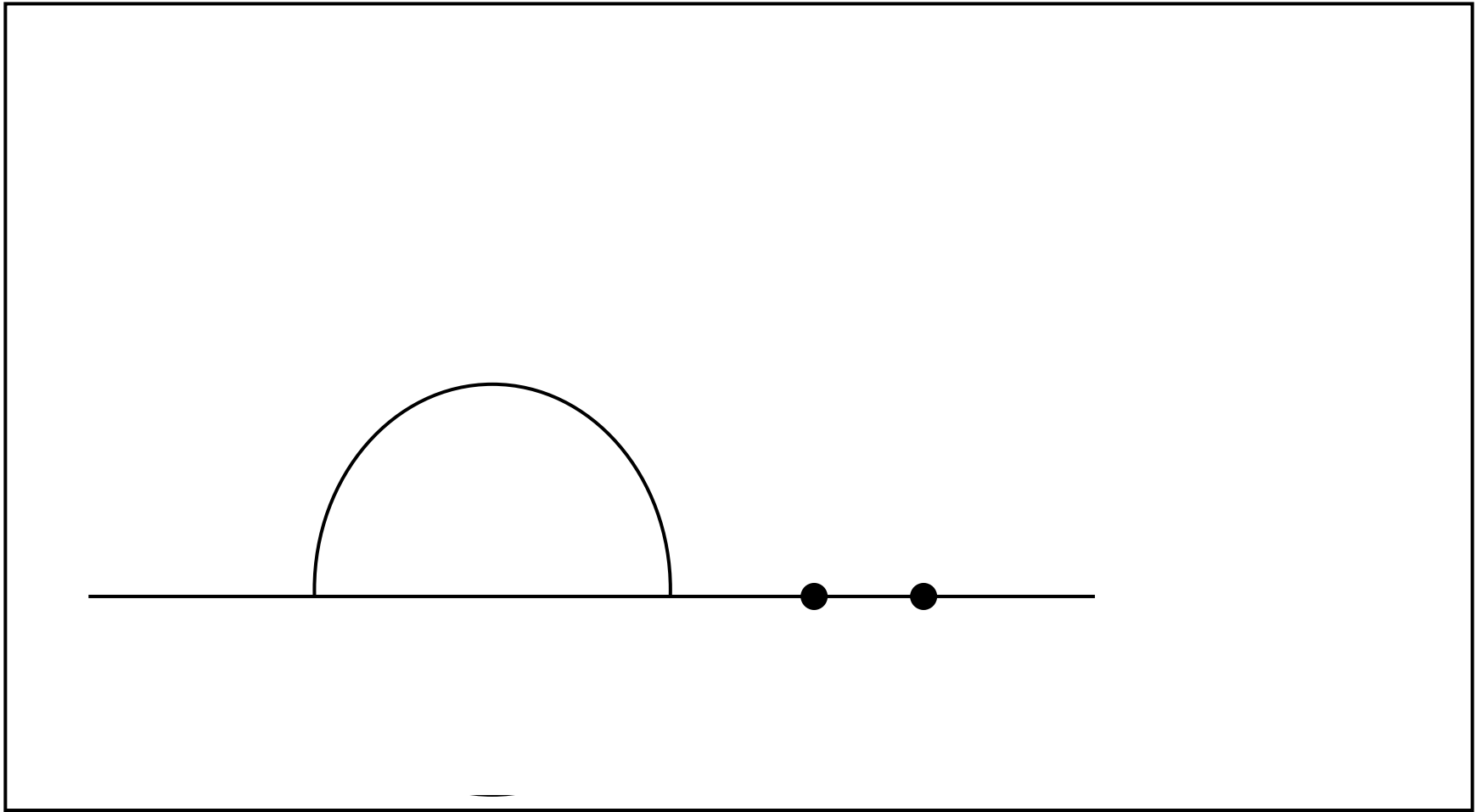
- We are in an exciting time of change.
- Information technology is a big driver of that change.
- The computer science theory of the last thirty years needs to be extended to cover the next thirty years.

Spectral Analysis

The model

$$G = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix} \rightarrow \hat{G} = \begin{pmatrix} \text{0-1} \\ \text{matrix} \end{pmatrix}$$

Eigenvalue distribution



Spectral Analysis

Recovering the graph structure

$$\hat{G} = UDU^T = \begin{pmatrix} u_1 & u_2 & \dots & u_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \lambda_3 & \\ & & & \lambda_4 & \dots \\ & & & & & \lambda_n \end{pmatrix} \begin{pmatrix} u_1^T & u_2^T & \dots & u_n^T \end{pmatrix}$$

$$\hat{G}(2) = \begin{pmatrix} \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & 0 & \\ & & & 0 & \dots \\ & & & & & 0 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

Power law distributions

$$M = DGD = \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{pmatrix} \rightarrow \hat{M}$$

If the degrees of a random matrix are power law distributed, the major eigenvectors are associated with neighborhoods of high degree vertices rather than structure of the graph.

Papadimitriou and Mihail

Signal to noise ratio

- Multiply every element of matrix by some fixed constant.
- The bounds in spectral analysis are determined by maximum noise of an element not the average.
- Multiplying low variance elements by constant increases signal without increasing maximum noise.

Variance of random variable

$$x = \begin{cases} 1 & p \\ 0 & 1-p \end{cases} \quad \sigma^2(x) = p(1-p) \cong p$$

Two ways of modifying x

$$y = \begin{cases} 1 & cp \\ 0 & 1 - cp \end{cases} \quad \sigma^2(y) = cp(1 - cp) \cong cp$$

$$z = \begin{cases} c & p \\ 0 & 1 - p \end{cases} \quad \sigma^2(z) = c^2 p(1 - p) \cong c^2 p$$

Increasing probability by c increases variance by c

Multiplying variable by c increases variance by c^2

Thus we correct for increase in probability of factor of c by multiplying variable by $\frac{1}{\sqrt{c}}$

$$L = \begin{pmatrix} \frac{1}{\sqrt{d_1}} & & & \\ & \frac{1}{\sqrt{d_2}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{d_n}} \end{pmatrix} \quad M = \begin{pmatrix} \frac{1}{\sqrt{d_1}} & & & \\ & \frac{1}{\sqrt{d_2}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{d_n}} \end{pmatrix}$$



- Power law distributions arise in many different contexts

- 1) data

- 2) queries

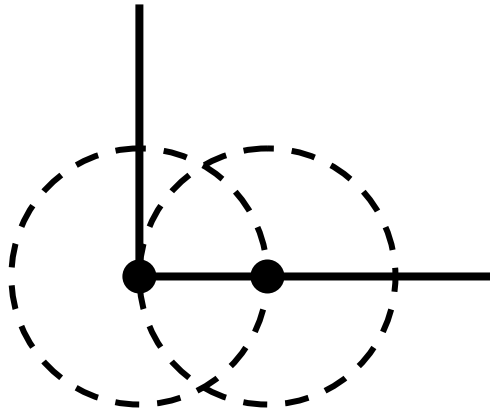
General theme emerging in clustering

- Use SVD to find reduced subspace
- Project data onto reduced subspace
- Cluster

Although SVD minimizes the sum of squared error between points and their expected values the error is not uniformly distributed and thus there are usually some outliers

- Reproject data onto subspace through the approximate cluster centers
- Recluster

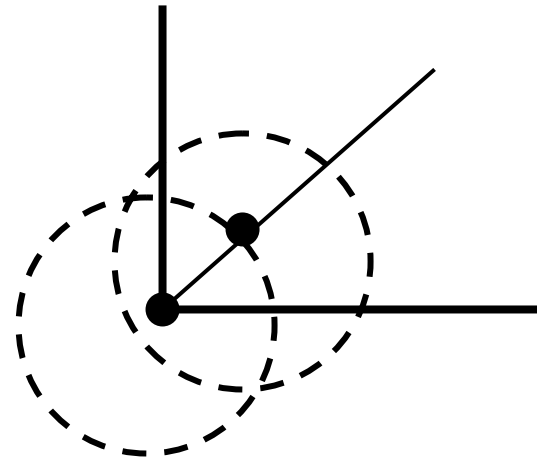
Two situations



Centers

$$(0, 0, \dots, 0)$$

$$(1, 0, \dots, 0)$$



Centers

$$(0, 0, \dots, 0)$$

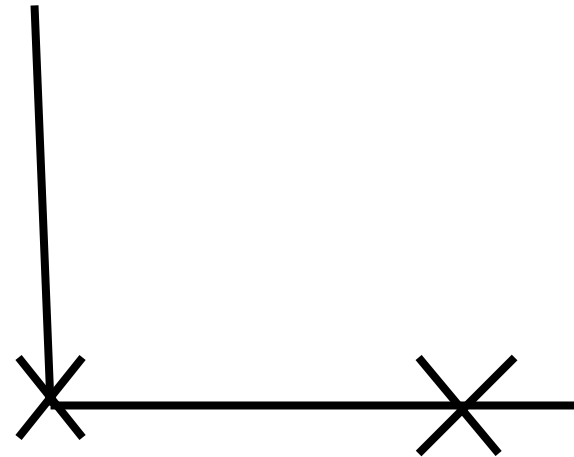
$$\frac{1}{\sqrt{d}}(1, 1, \dots, 1)$$

Two situations

- Centers

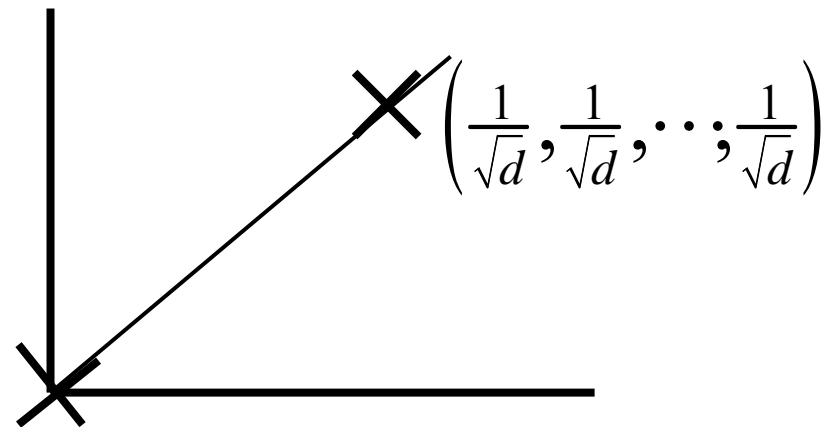
$$(0, 0, \dots, 0)$$


$$(1, 0, \dots, 0)$$



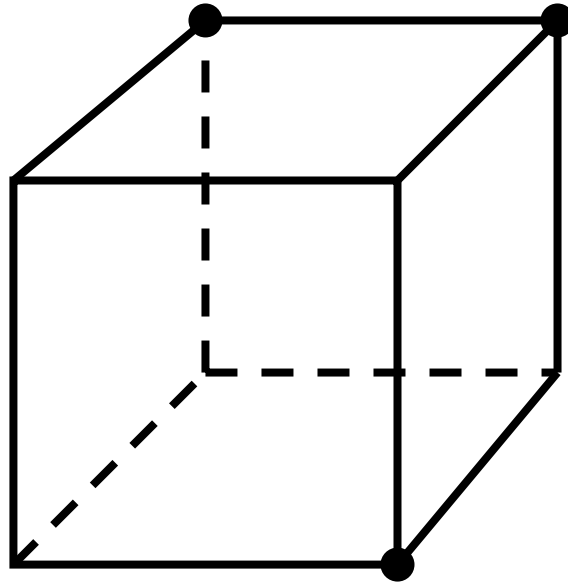
- Centers $(0, 0, \dots, 0)$

$$\left(\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}, \dots, \frac{1}{\sqrt{d}}\right)$$



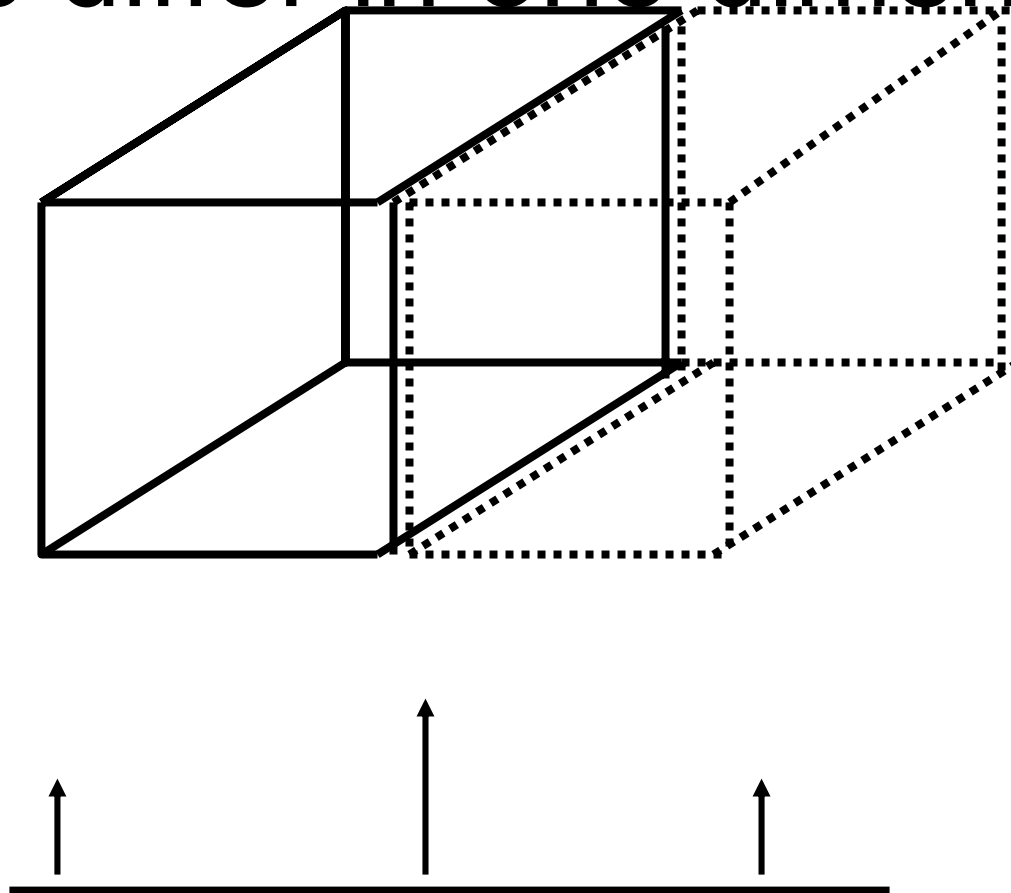
- 
- For spherical Gaussian the two situations are equivalent – Probability distribution for spherical Gaussian depends only on distance from center.
 - For binomial distributions – the two situations are fundamentally different.

Binomial distribution in d dimensions

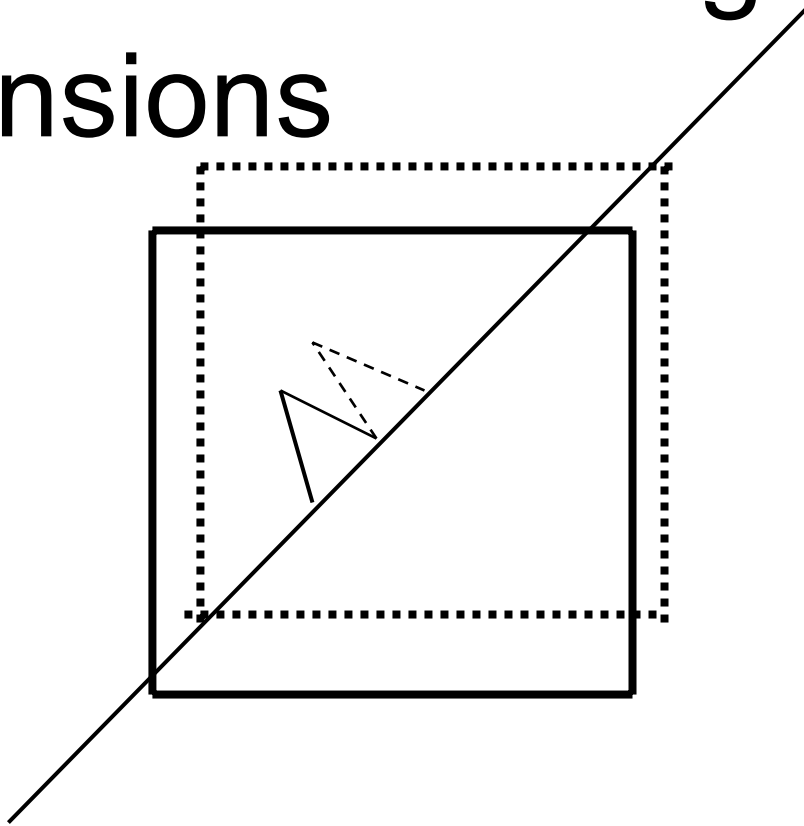


$$x_i = \begin{cases} 0 & p = \frac{1}{2} \\ 1 & p = \frac{1}{2} \end{cases}$$

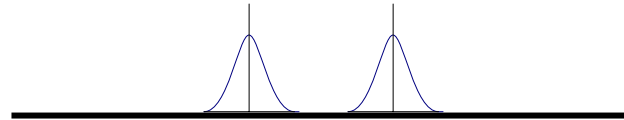
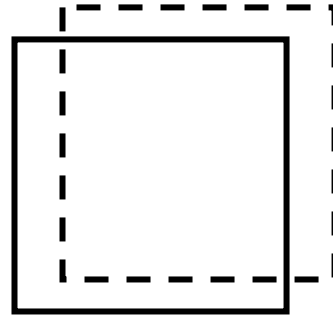
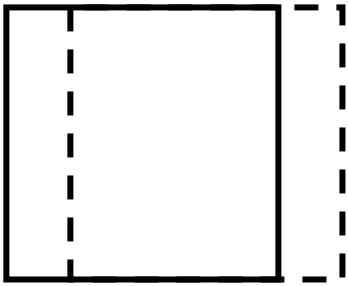
Centers differ in one dimension



Centers differ along all dimensions



The two situations



Balance

- Suggests we define balance of a unit vector by how uniformly the coordinates contribute to its length.

$$\text{bal}(x) = \frac{|x|_{\infty}}{|x|_2}$$

General method

- Project data onto SVD subspace
- Cluster
- Draw lines through all k^2 pairs of cluster centers
- Smooth lines
- Project onto each line and cluster

