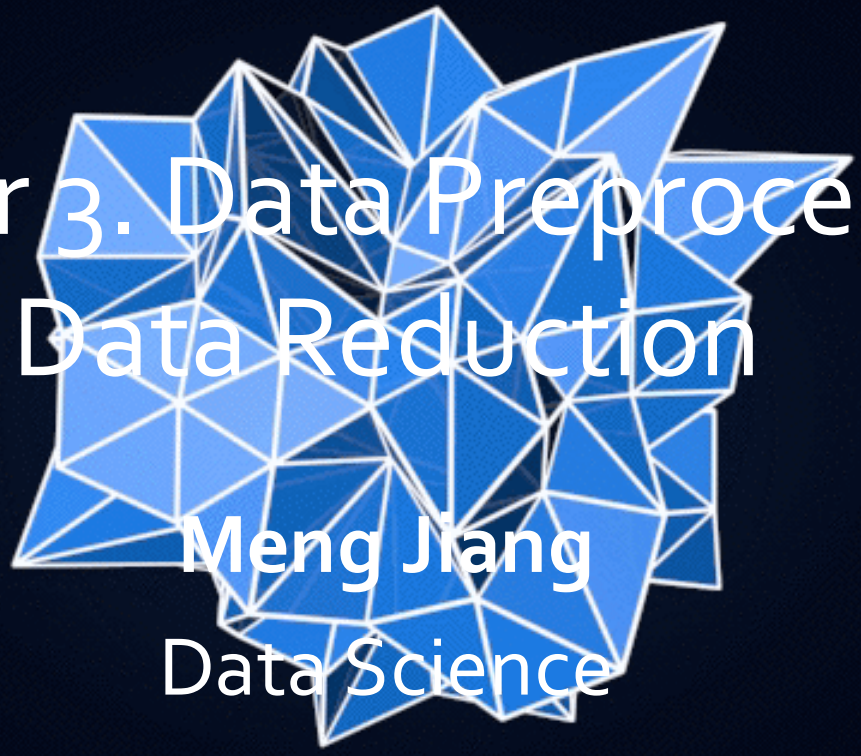
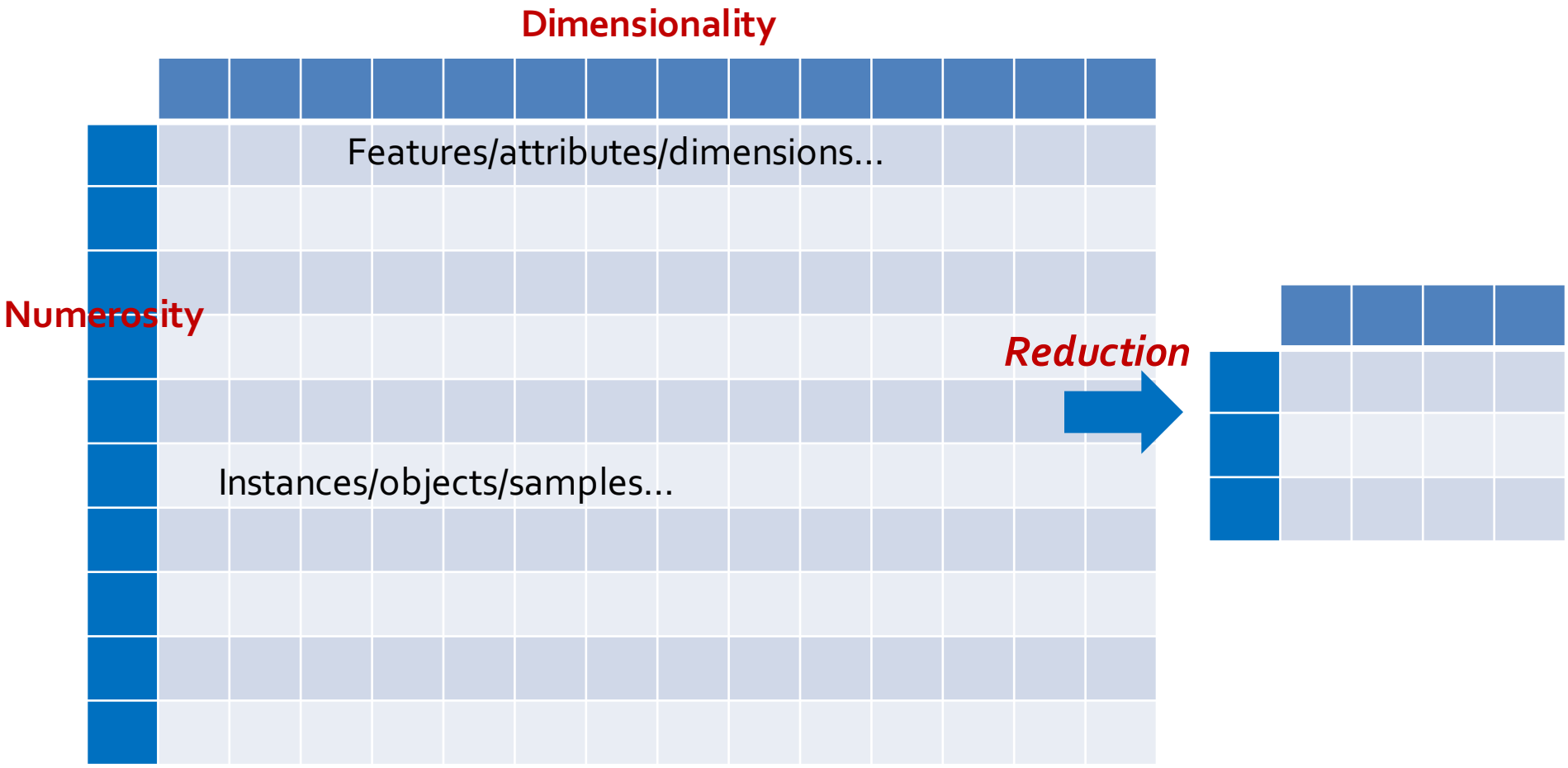
Chapter 3. Data Preprocessing: Data Reduction



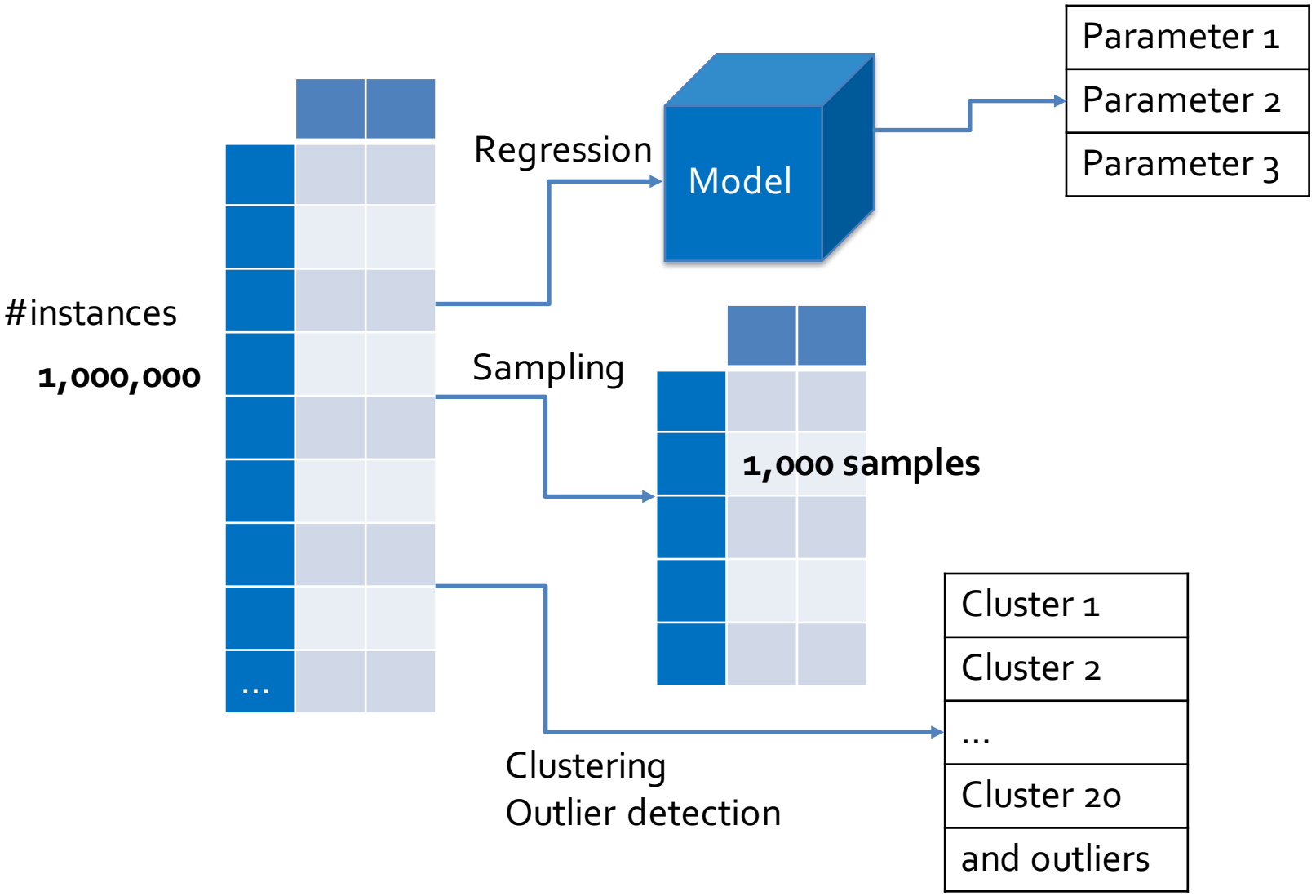
Meng Jiang
Data Science

error23

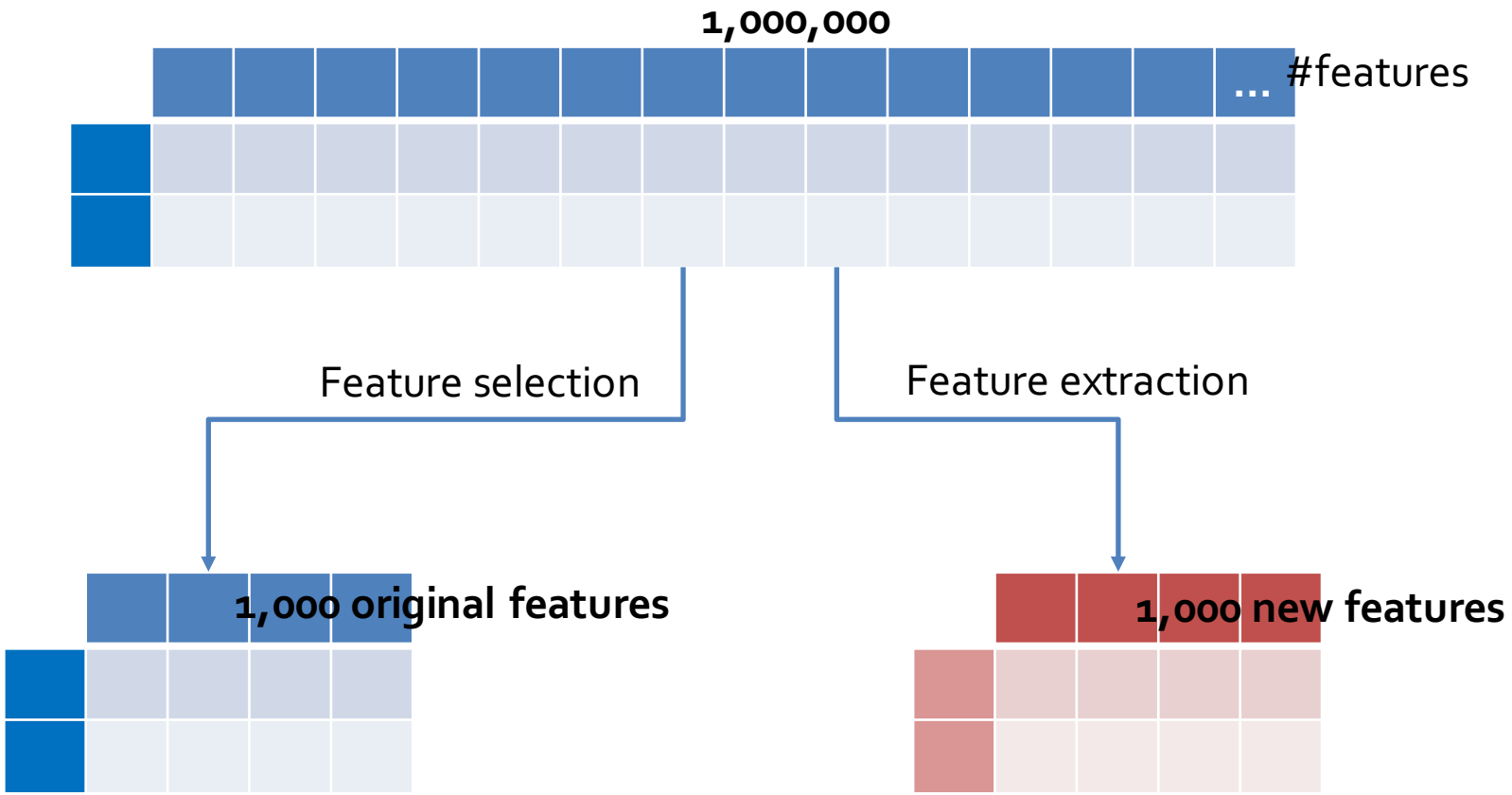
Numerosity and Dimensionality



Numerosity Reduction



Dimensionality Reduction

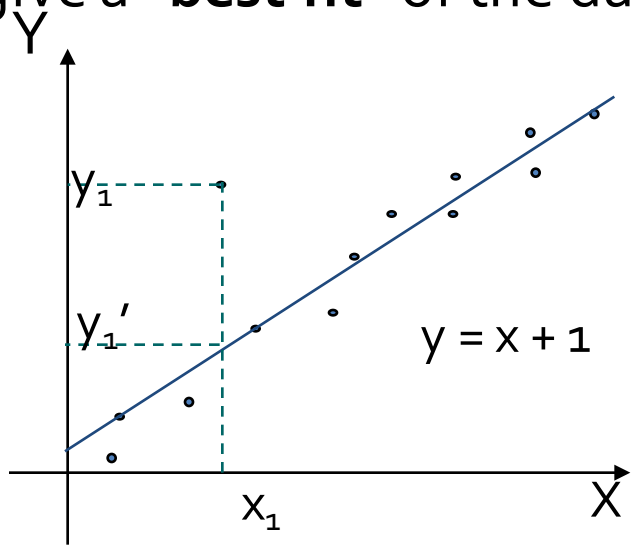


Today: Data Reduction

- Describe **numerosity reduction** (reducing #instances)
  - Parametric methods: Fit some model and estimate model parameters
    - Regression: Describe linear/non-linear regression models
  - Nonparametric methods
    - Histograms
    - Clustering
    - Sampling: Describe stratified sampling
- Describe **dimensionality reduction** (reducing #features)
  - Feature selection
    - Heuristic search
  - Feature extraction
    - Principal component analysis (PCA)
    - Singular value decomposition (SVD)

Regression Analysis

- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values
  - of a ***dependent variable*** (also called ***response variable*** or ***measurement***): Y
  - and of one or more *independent variables* (also known as ***explanatory variables*** or ***predictors***): X, or X_1, X_2, …X_n
- Parameters are estimated to give a "**best fit**" of the data
  - Data: (x_1, y_1)
  - Fit of the data: (x_1, y_1')
    - Ex. $y_1' = x_1 + 1$

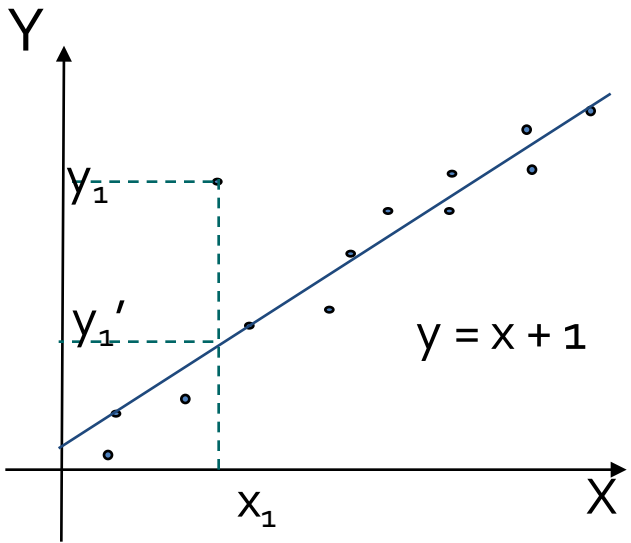


Least Square Method for Regresssion

- Most commonly the best fit is evaluated by using the ***least square method***, but other criteria have also been used

  $\min g = \sum_{i=1}^{n}(y_i - y'_i)^2$, where $y'_i = f(x_i, \beta)$

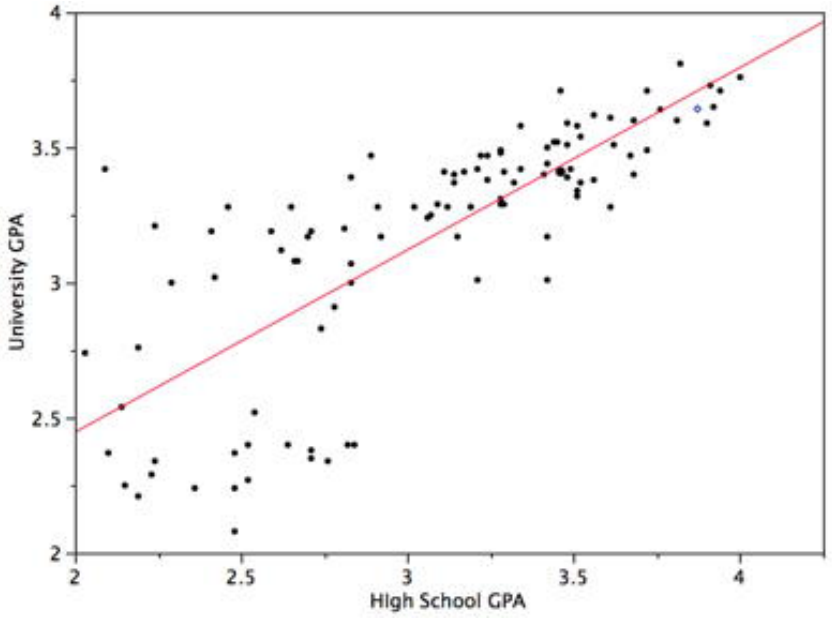
- Used for **prediction** (including forecasting of time-series data), **inference**, and **hypothesis testing**.



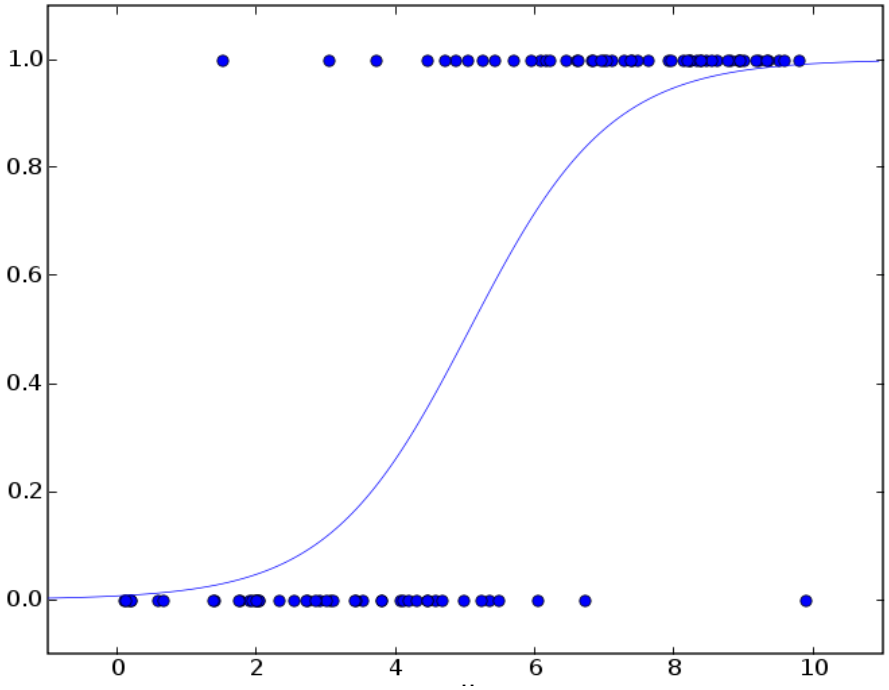
Set up y = f(x) = β_1 x + β_2
Learn β by minimizing the least square error

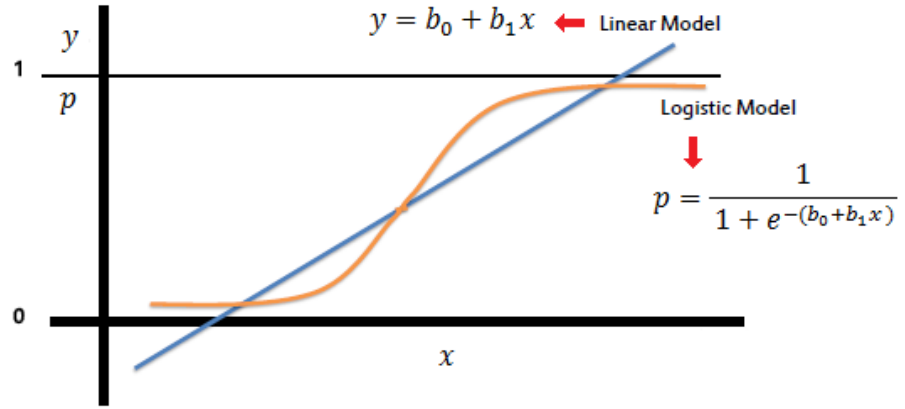
Linear Regression

- Linear regression: $Y = w X + b$
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
  - Two regression coefficients, w and b, specify the line and are to be estimated by using the data at hand



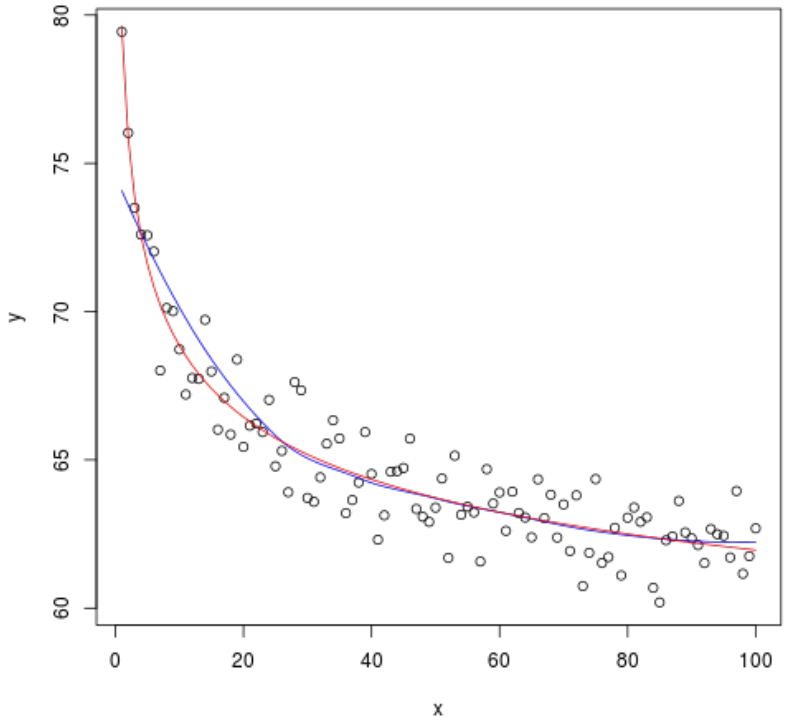
Logistic Regression





Nonlinear Regression

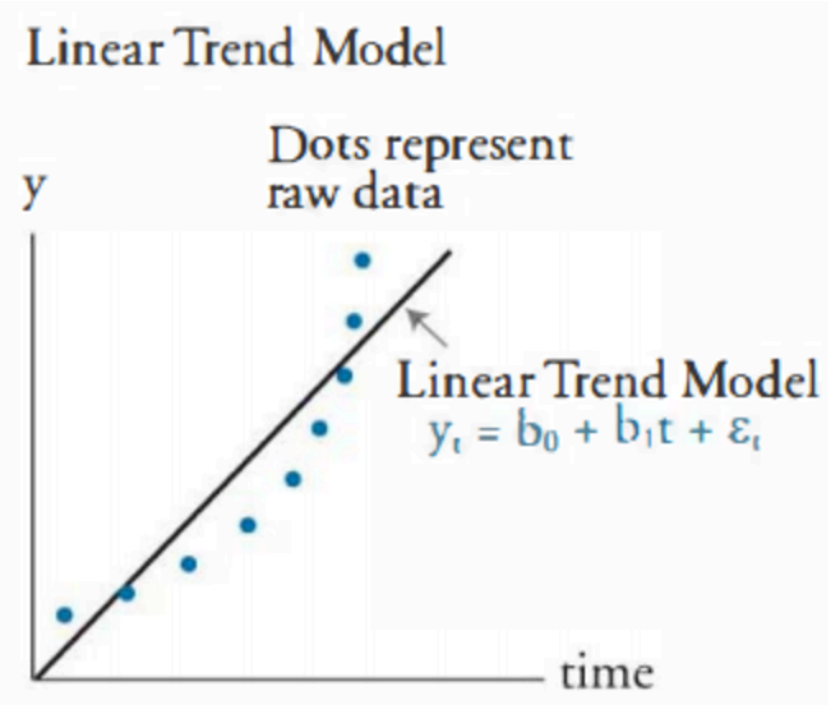
- Nonlinear regression:
  - Data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables

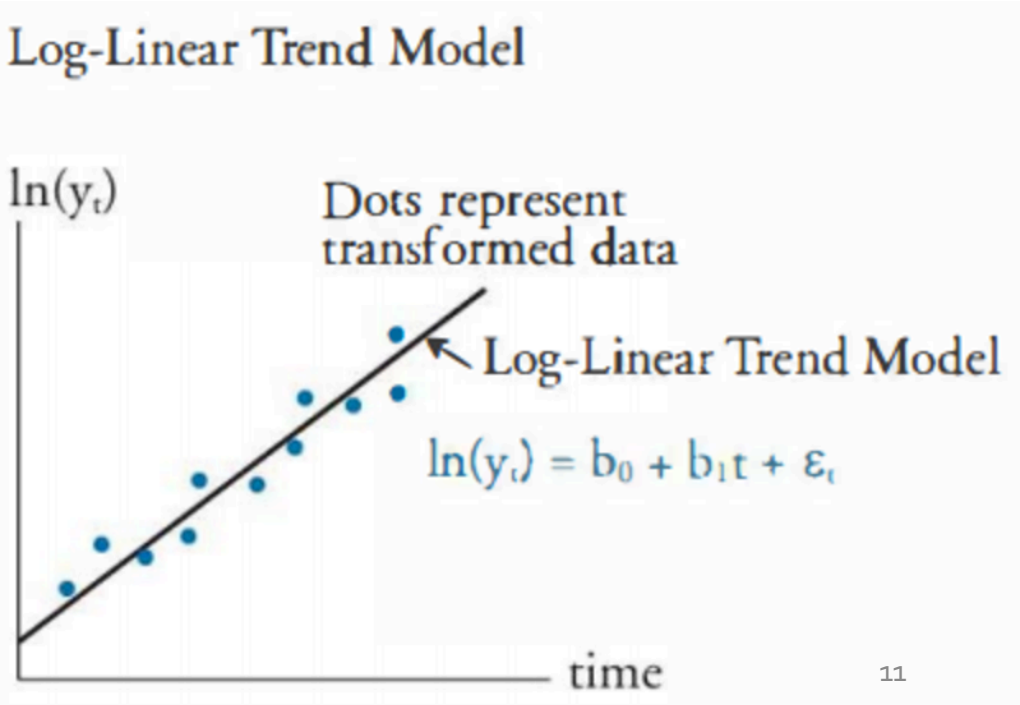


Log-Linear Model

- Log-linear model
  - A math model that takes the form of a **function whose logarithm** is a **linear** combination of the parameters of the model

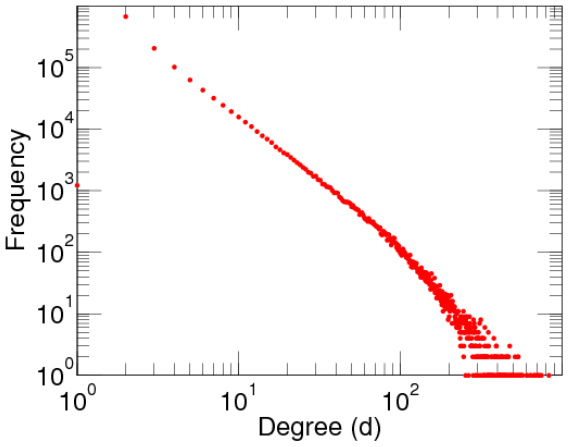
Q: How about Log-Log model?



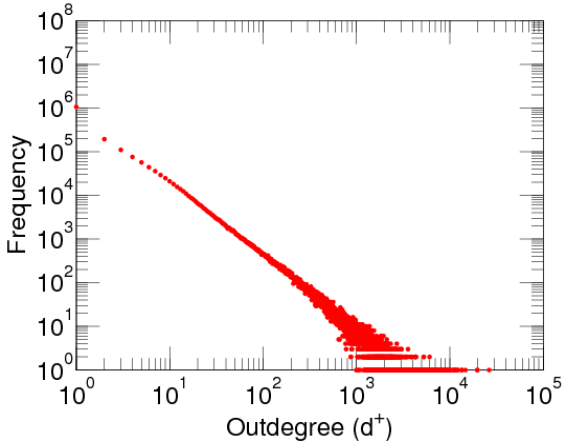


Log-Log: Power Law

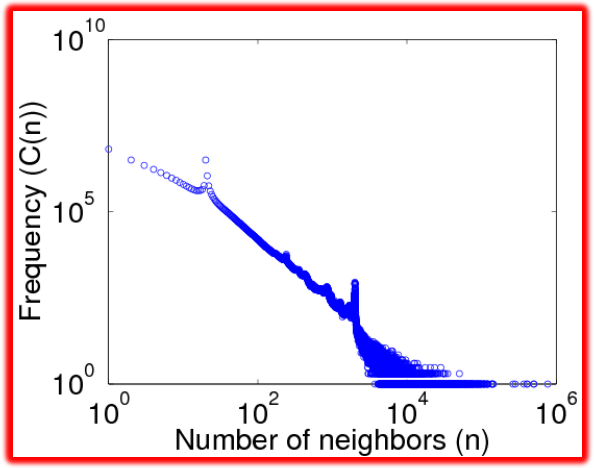
DBLP
Author-publication



Flickr
User-user



Twitter
Who-follows-whom

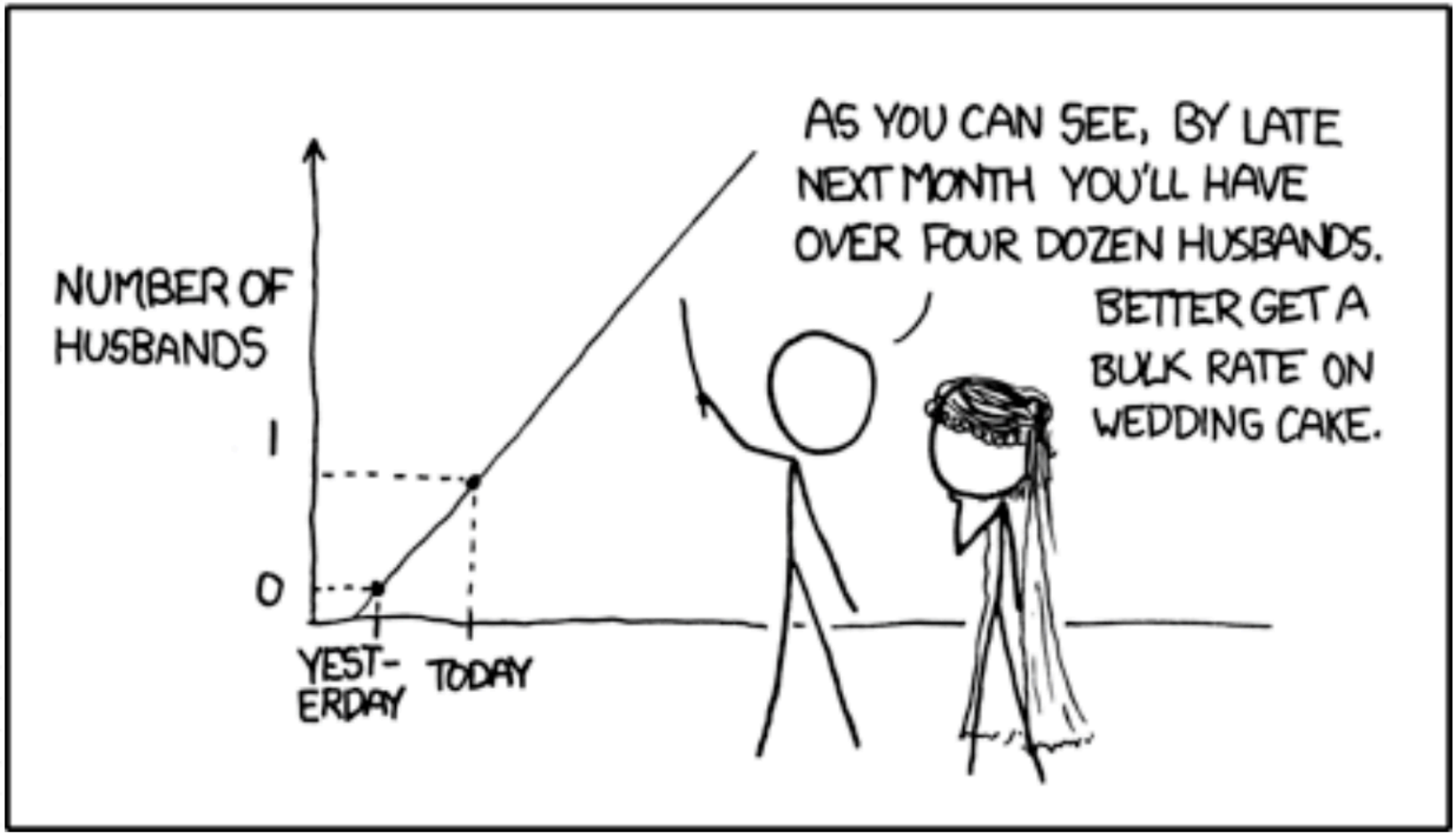


[konect.uni-koblenz.de/networks/]

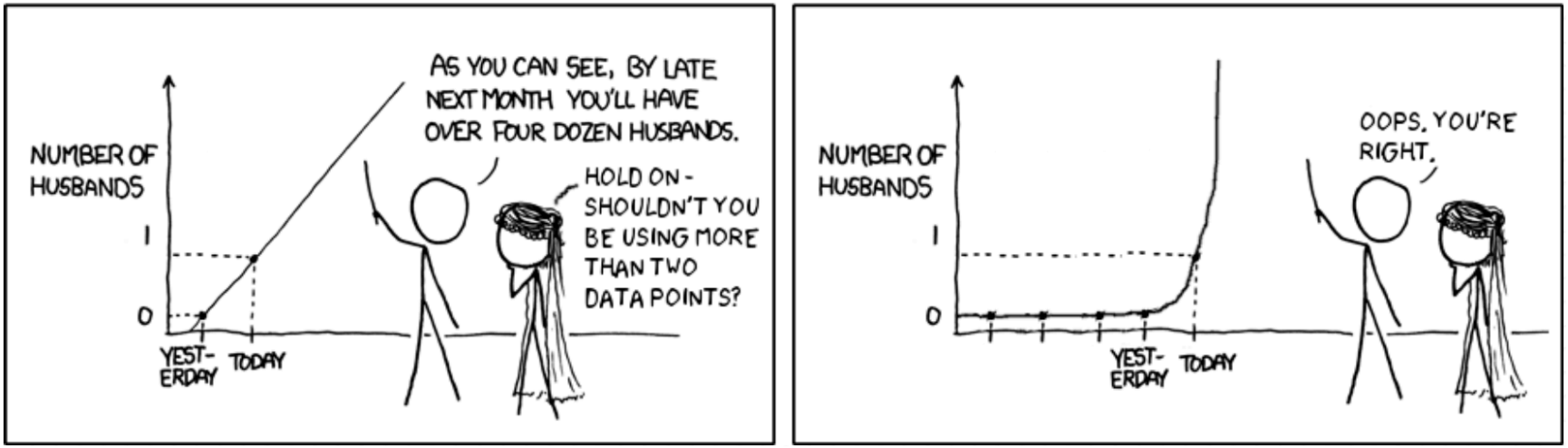
Out-degree distributions (network):

Power-law distributions in networks [Faloutsos et al. SIGCOMM'99; Chung et al. PNAS'02]

Caution: Extrapolation



Caution: Extrapolation

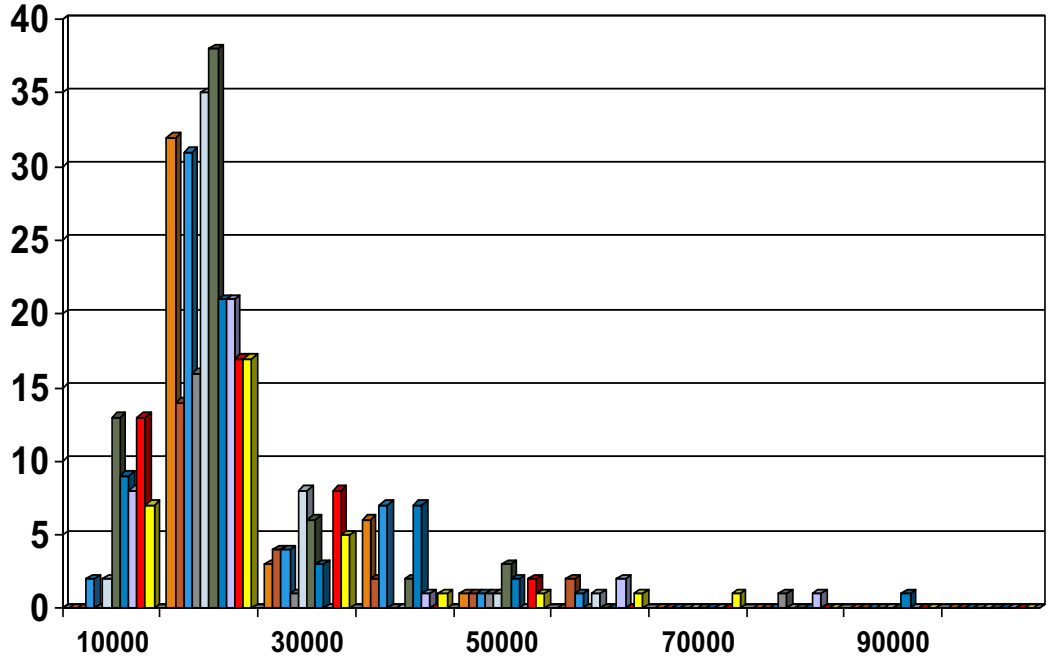


Today: Data Reduction

- Describe **numerosity reduction** (reducing #instances)
  - Parametric methods: Fit some model and estimate model parameters
    - Regression: Describe linear/non-linear regression models
  - Nonparametric methods
    - Histograms
    - Clustering
    - Sampling: Describe stratified sampling
- Describe **dimensionality reduction** (reducing #features)
  - Feature selection
    - Heuristic search
  - Feature extraction
    - Principal component analysis (PCA)
    - Singular value decomposition (SVD)

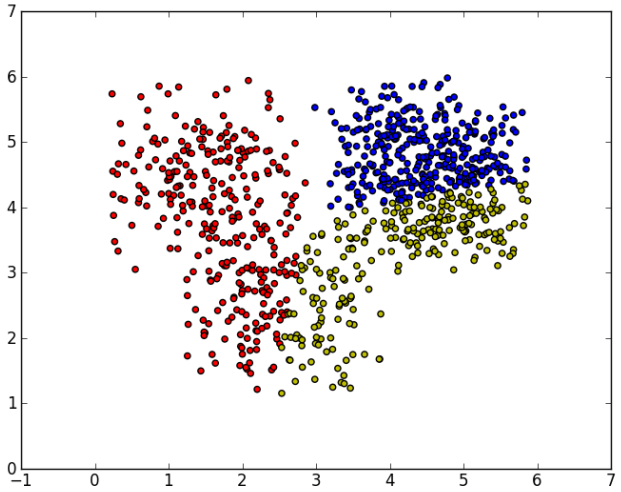
Histograms

- Divide data into buckets and store average (sum) for each bucket



Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., **centroid and diameter**) only
- Can be very effective if data is clustered but not if data is "smeared"
- There are many choices of clustering definitions and clustering algorithms



Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Key principle: Choose a **representative** subset of the data
  - Simple random sampling may have very poor performance in the presence of skew

Simple random sampling:

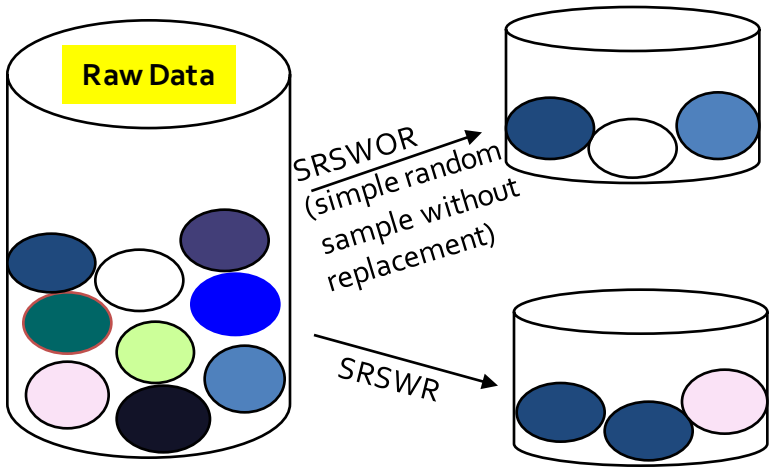
Equal probability of selecting any particular item

Sampling without replacement:

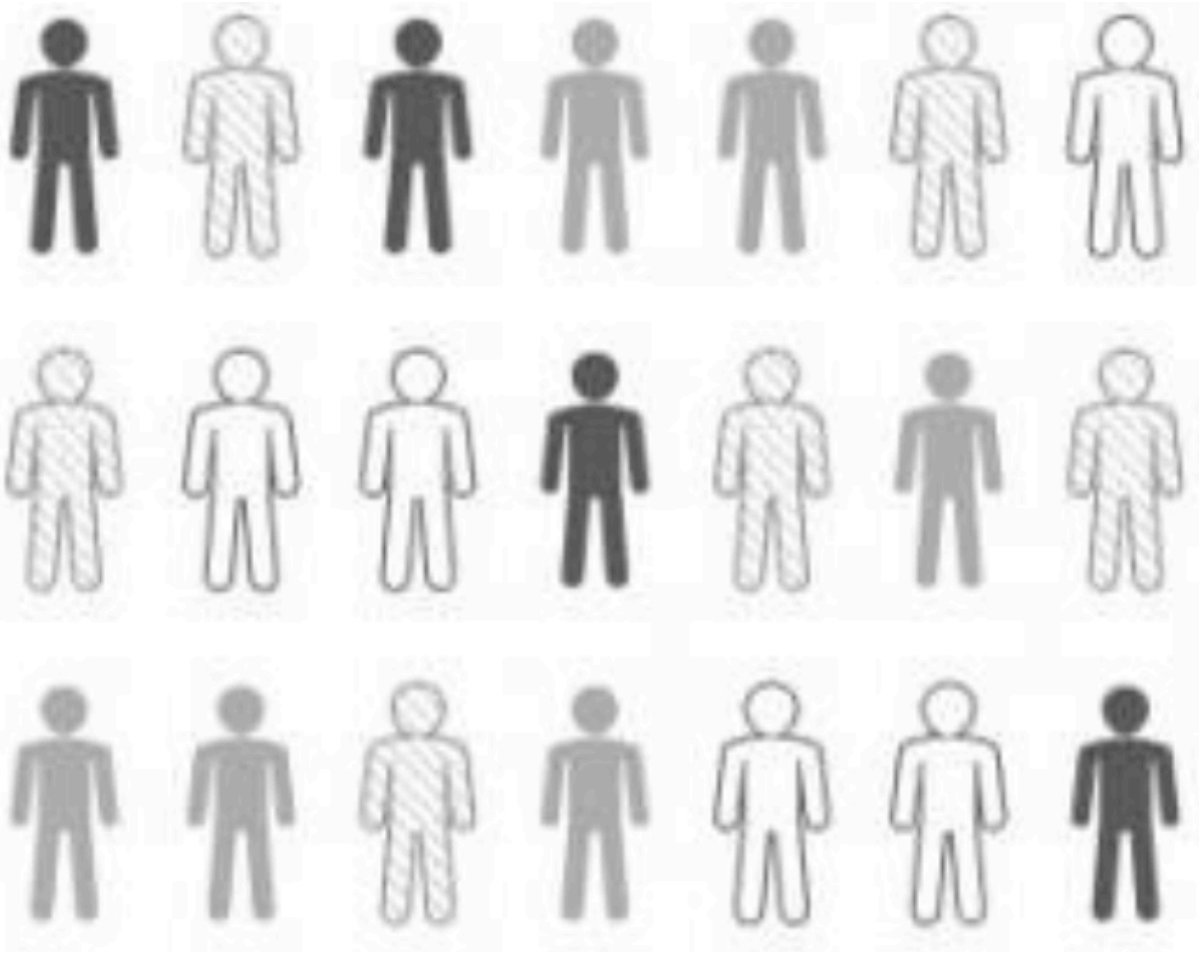
Once an object is selected, it is removed from the population

Sampling with replacement:

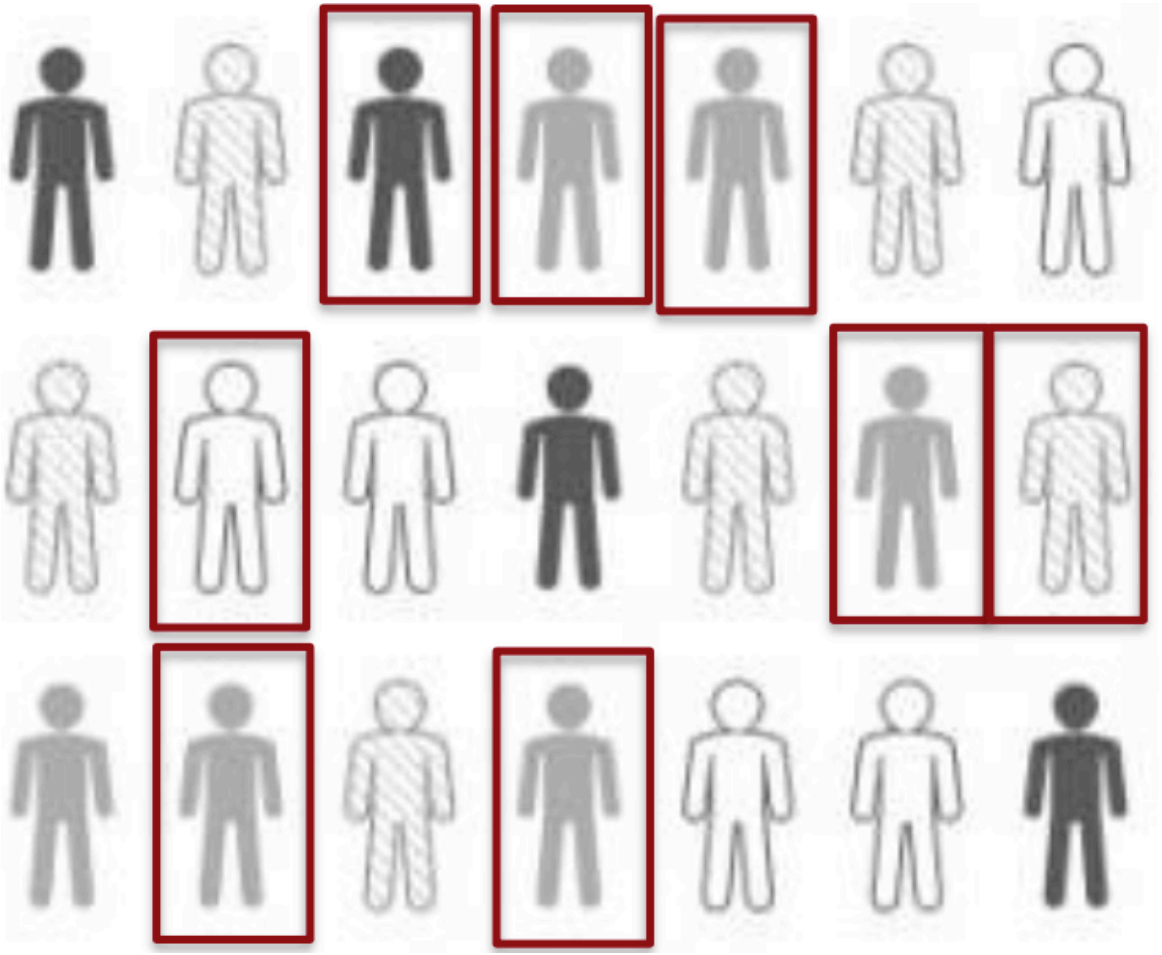
A selected object is not removed from the population



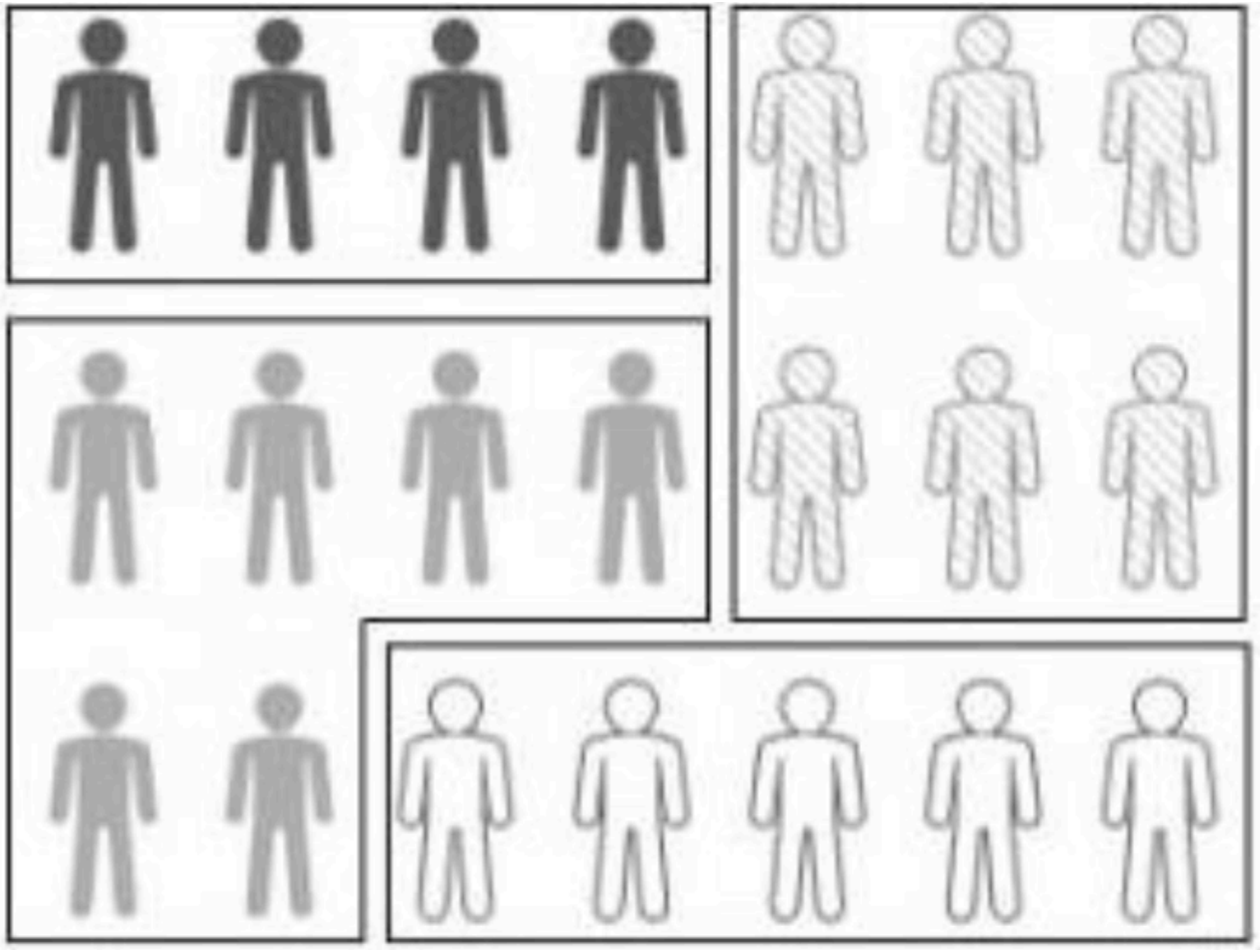
What if the data is ***imbalanced***?



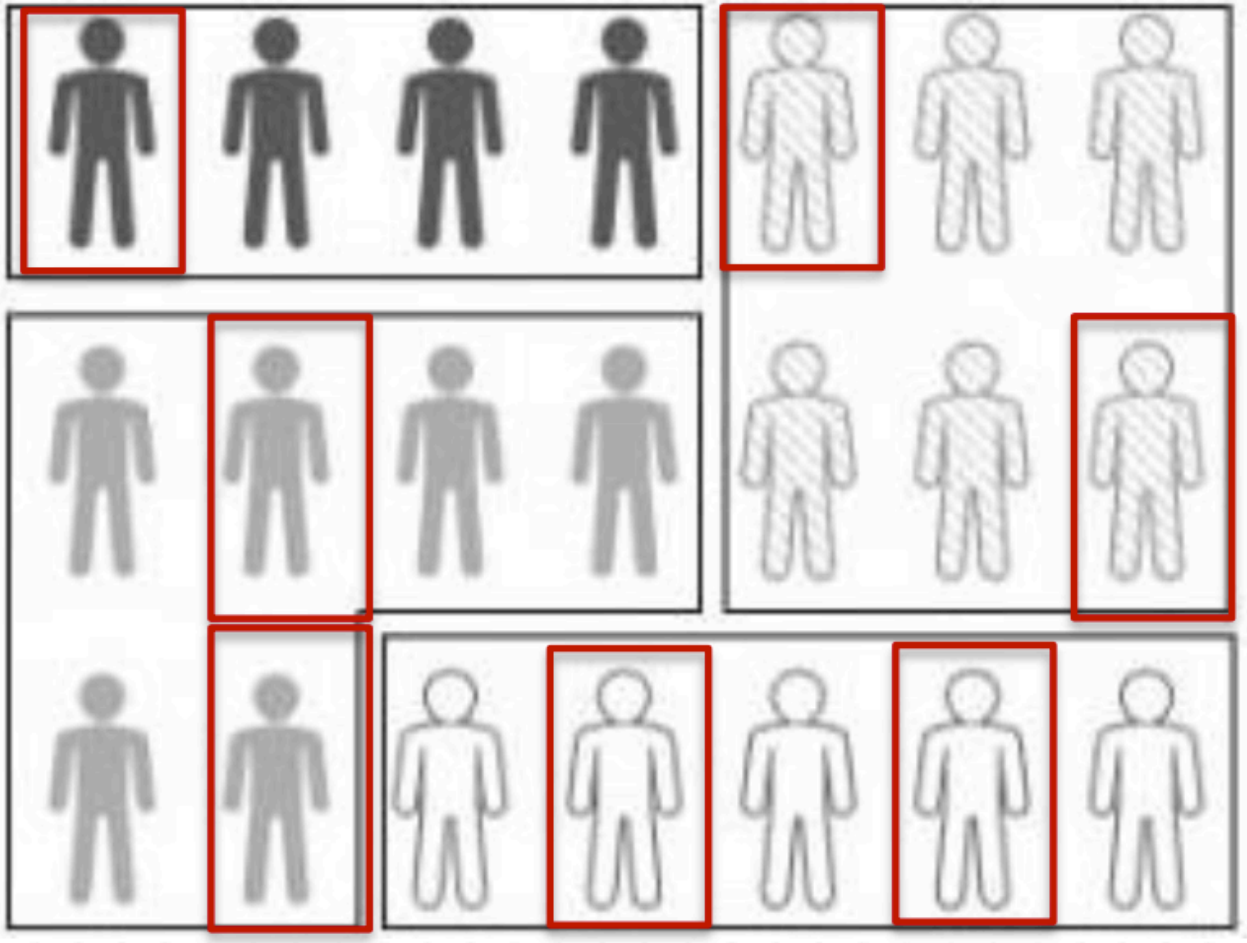
What if the data is ***imbalanced***?



What if the data is ***imbalanced***?

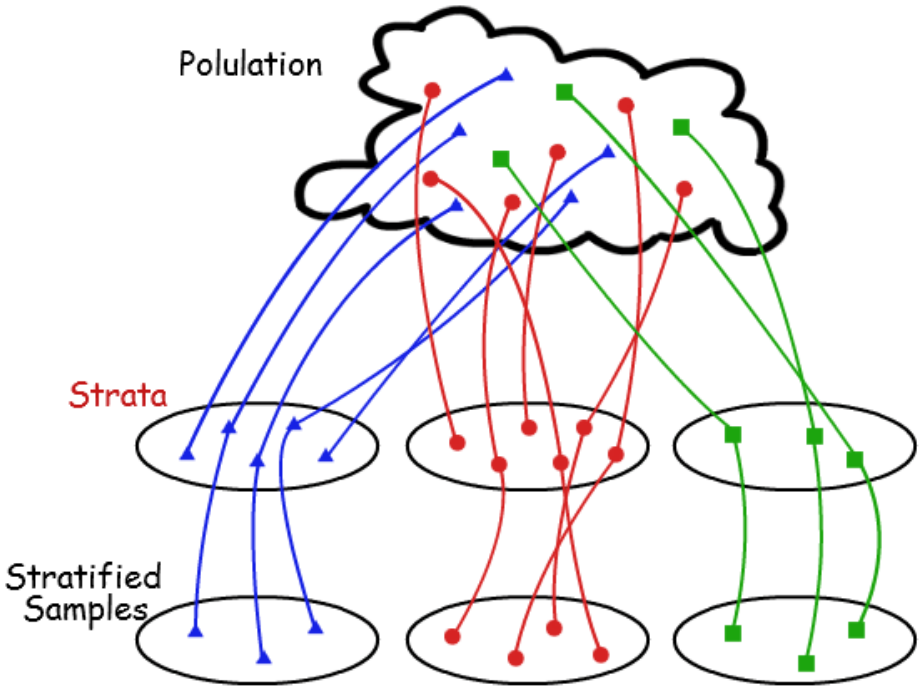


What if the data is ***imbalanced***?



Stratified Sampling

- **Stratified sampling**
  - Partition (or cluster) the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)



Today: Data Reduction

- Describe **numerosity reduction** (reducing #instances)
  - Parametric methods: Fit some model and estimate model parameters
    - Regression: Describe linear/non-linear regression models
  - Nonparametric methods
    - Histograms
    - Clustering
    - Sampling: Describe stratified sampling
- Describe **dimensionality reduction** (reducing #features)
  - Feature selection
    - Heuristic search
  - Feature extraction
    - Principal component analysis (PCA)
    - Singular Value Decomposition (SVD)

Dimensionality Reduction Methodologies

- **Feature selection**: Find a subset of the original variables (or features, attributes)
  - Heuristic search in attribute subset selection
  - …
- **Feature extraction**: Transform the data in the high-dimensional space to a space of fewer dimensions
  - Principal component analysis
  - …

Attribute Subset Selection

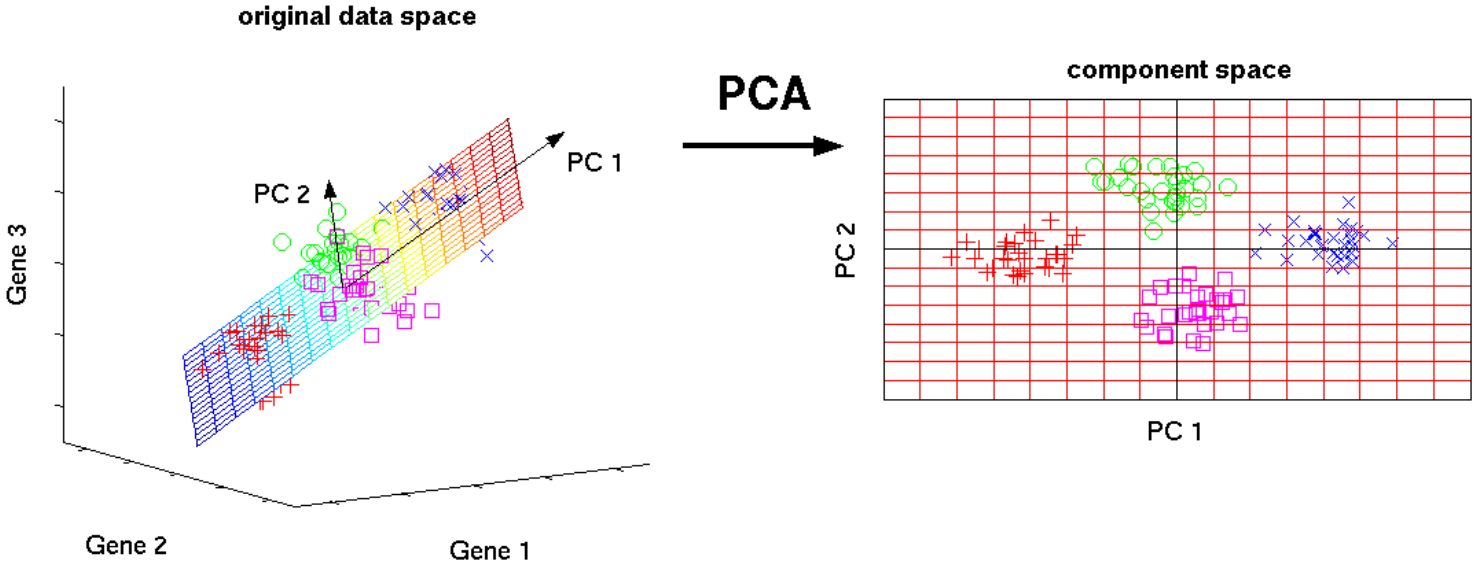
- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
    - Purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
    - A student's ID is often irrelevant to the task of predicting his/her GPA

Heuristic Search in Attribute Selection

- There are 2^d possible attribute combinations of d attributes
- Typical heuristic attribute selection methods:
  - **Method 1:** Best single attribute under the attribute independence assumption: choose by significance tests
  - **Method 2:** Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - **Method 3:** Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - **Method 4:** Best combined attribute selection and elimination

Principal Component Analysis (PCA)

- PCA: A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called ***principal components***
- The original data are projected onto a **much smaller space**, resulting in dimensionality reduction (e.g., n=3 to k=2)

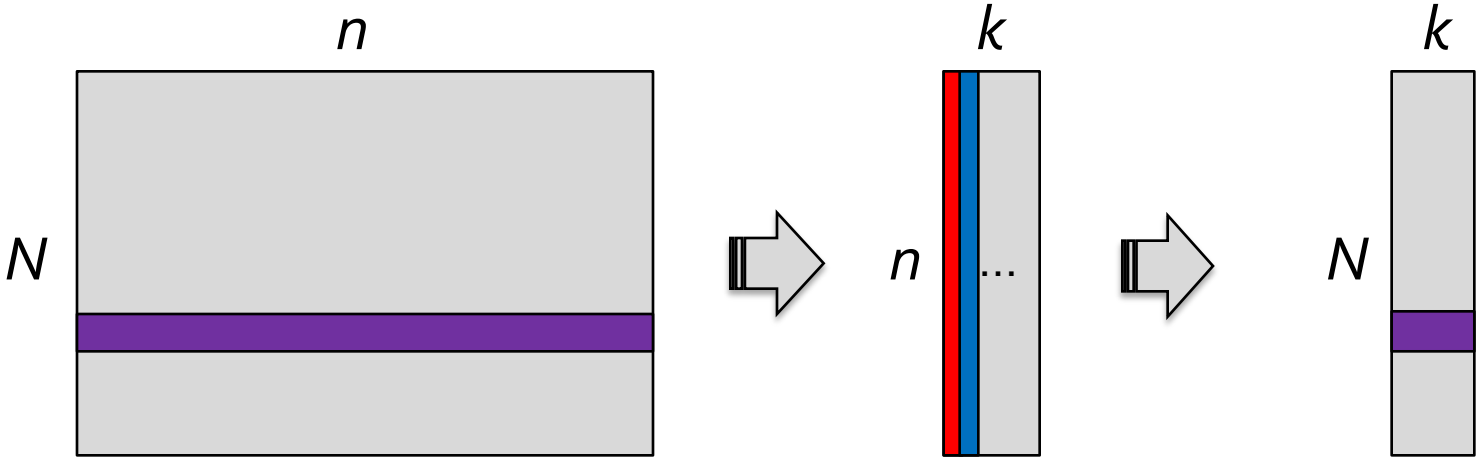


PCA (cont.)

- **Given** N data vectors from n-dimensions, **find** k ≤ n orthogonal vectors (principal components) best used to represent data
  - **Normalize** input data: Each attribute falls within the same range
  - Compute k **orthonormal (unit) vectors**, i.e., principal components

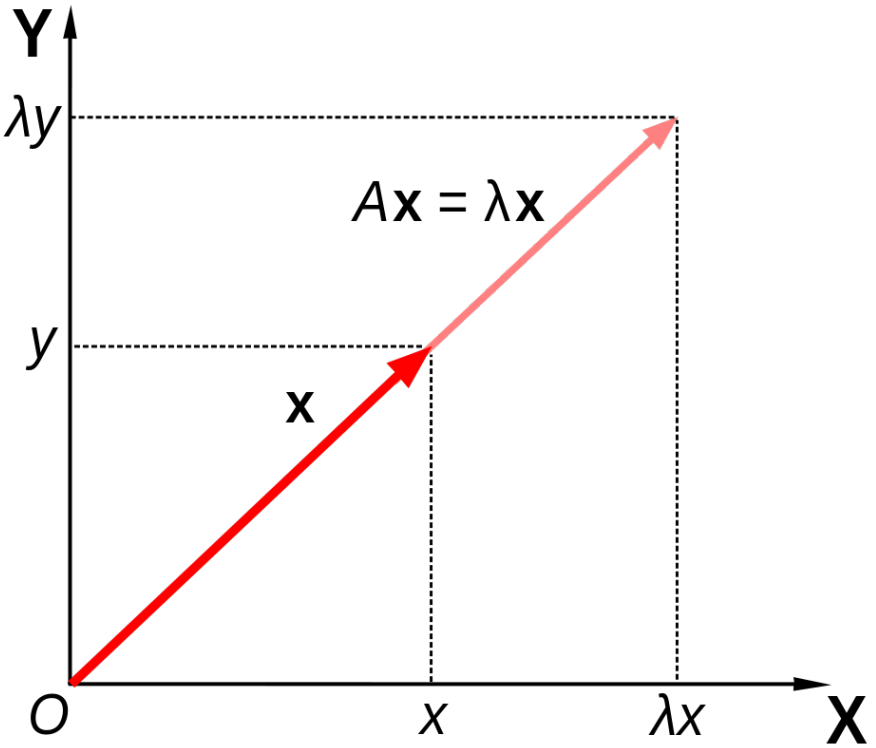
normalized eigenvector

- Each input data (vector) is a linear combination of the k **principal components**

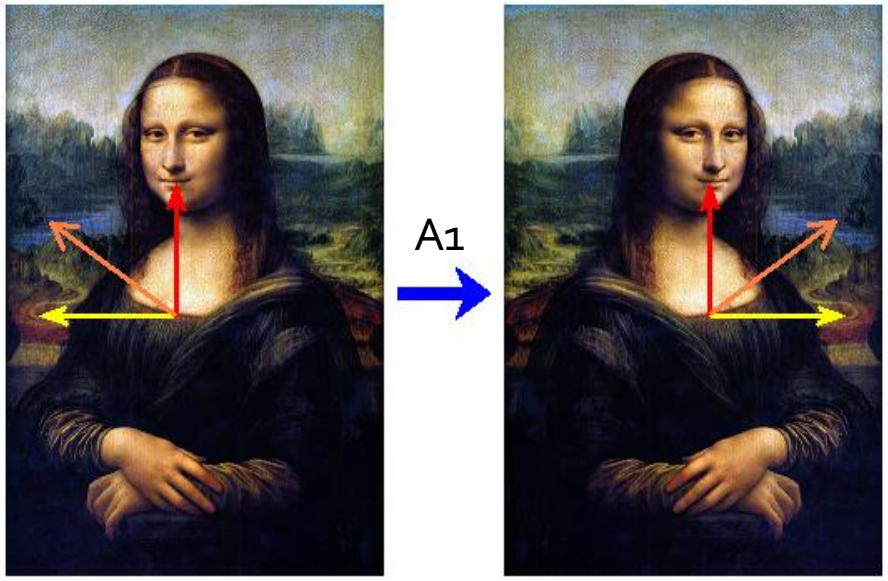


Eigenvectors (cont.)

- For a square matrix **A** (n*n), find the eigenvector **x** (n*1).
  - **A** represents the linear transformation (from n to n)
- Matrix **A** acts by stretching the vector **x**, not changing its direction, so **x** is an eigenvector of **A**.



Eigenvectors (cont.)



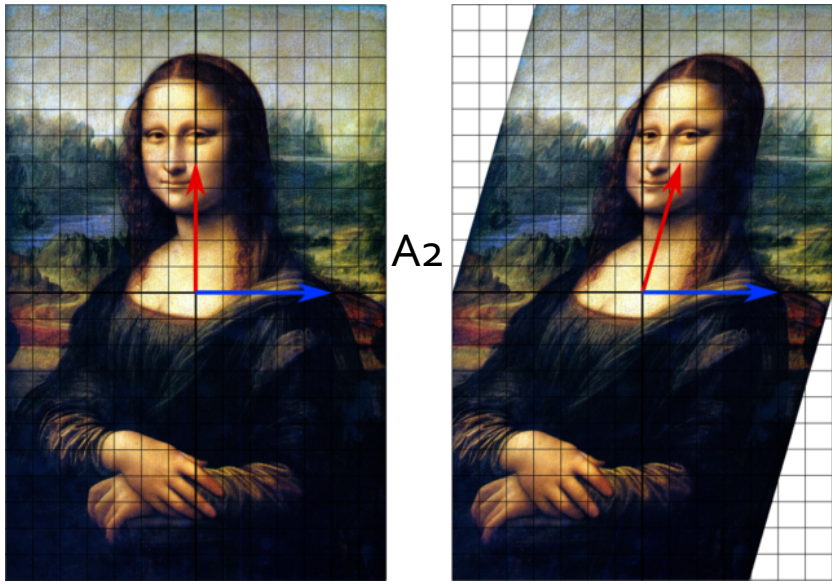
Which vectors are eigenvectors?

- Red
- Orange
- Yellow

What are the eigenvalues?

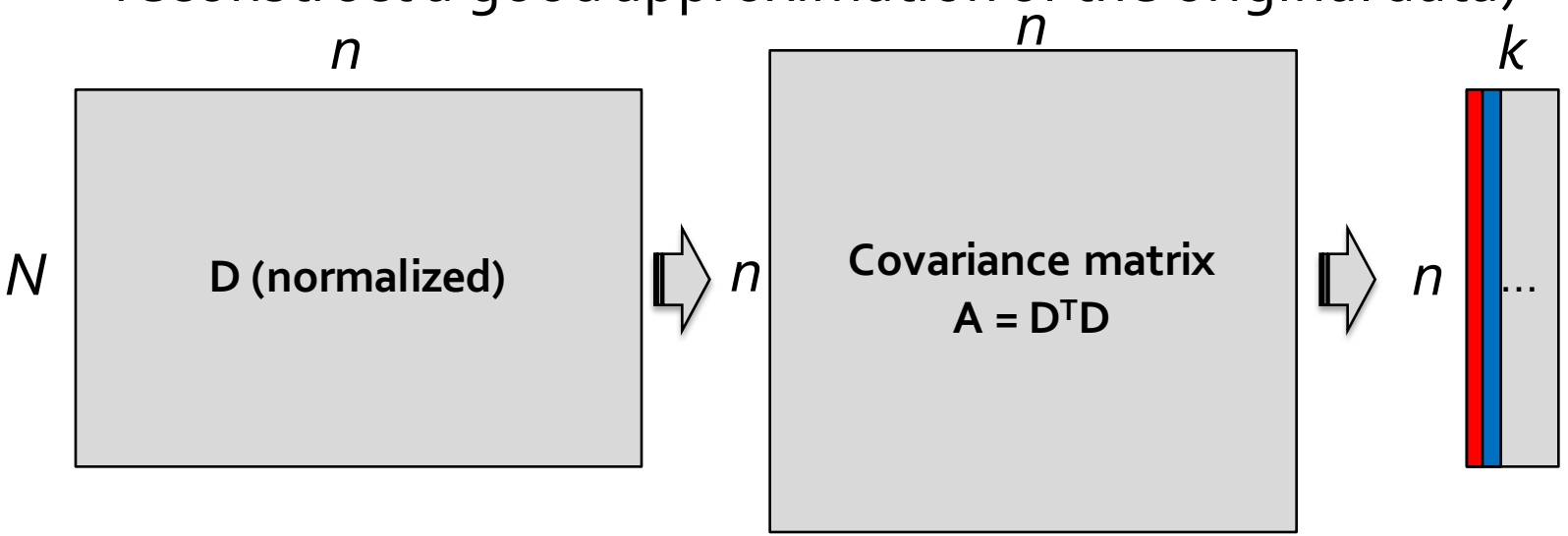
Which vectors are eigenvectors?

- Red
- Blue



PCA and Eigenvectors

- For ***Square Matrix***: Data matrix to Covariance matrix
- The principal components are sorted in order of **decreasing "significance" or strength**
- **From n to k:** Since the components are sorted, the size of the data can be reduced by eliminating the weak components (i.e., using the strongest principal components, to reconstruct a good approximation of the original data)



PCA and Eigenvectors (cont.)

- Method: Find the **eigenvectors of covariance (square) matrix**, and these eigenvectors define the new space

$$\begin{aligned} \mathbf{Ax} = \lambda\mathbf{x} &\Leftrightarrow \mathbf{Ax} - \lambda\mathbf{x} = \mathbf{0} \\ &\Leftrightarrow \mathbf{Ax} - \lambda\mathbf{Ix} = \mathbf{0} \\ &\Leftrightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \end{aligned}$$

The equation $\mathbf{Ax} = \lambda\mathbf{x}$ has nonzero solutions for the vector x if and only if the matrix $\mathbf{A} - \lambda\mathbf{I}$ has zero determinant.

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

Ex. Eigenvalues

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

The eigenvalues are those λ for which $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. Now

$$\begin{aligned}
\det(\mathbf{A} - \lambda\mathbf{I}) &= \det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \\
&= \det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) \\
&= \begin{vmatrix} 2-\lambda & 2 \\ 5 & -1-\lambda \end{vmatrix} \\
&= (2-\lambda)(-1-\lambda) - 10 \\
&= \lambda^2 - \lambda - 12.
\end{aligned}$$

The eigenvalues of $\mathbf{A}$ are the solutions of the quadratic equation $\lambda^2 - \lambda - 12 = 0$, namely $\lambda_1 = -3$ and $\lambda_2 = 4$.

Ex. Eigenvectors

First, we work with $\lambda = -3$. The equation $\mathbf{Ax} = \lambda\mathbf{x}$ becomes $\mathbf{Ax} = -3\mathbf{x}$. Writing

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and using the matrix $\mathbf{A}$ from above, we have

$$\mathbf{Ax} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix},$$

while

$$-3\mathbf{x} = \begin{bmatrix} -3x_1 \\ -3x_2 \end{bmatrix}.$$

Setting these equal, we get

$$\begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix} = \begin{bmatrix} -3x_1 \\ -3x_2 \end{bmatrix} \Rightarrow \quad 2x_1 + 2x_2 = -3x_1 \quad \text{and} \quad 5x_1 - x_2 = -3x_2$$

$$\Rightarrow \quad 5x_1 = -2x_2$$

$$\Rightarrow \quad x_1 = -\frac{2}{5}x_2.$$

$$\mathbf{u_1} = \begin{bmatrix} 2 \\ -5 \end{bmatrix}$$

Ex. Eigenvectors (cont.)

Similarly, we can find eigenvectors associated with the eigenvalue $\lambda = 4$ by solving $\mathbf{Ax} = 4\mathbf{x}$:

$$\begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix} = \begin{bmatrix} 4x_1 \\ 4x_2 \end{bmatrix} \Rightarrow 2x_1 + 2x_2 = 4x_1 \quad \text{and} \quad 5x_1 - x_2 = 4x_2$$
$$\Rightarrow x_1 = x_2.$$

Hence the set of eigenvectors associated with $\lambda = 4$ is spanned by

$$\mathbf{u_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Ex. Eigenvalues (cont.)

Example: Find the eigenvalues and associated eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} 7 & 0 & -3 \\ -9 & -2 & 3 \\ 18 & 0 & -8 \end{bmatrix}.$$

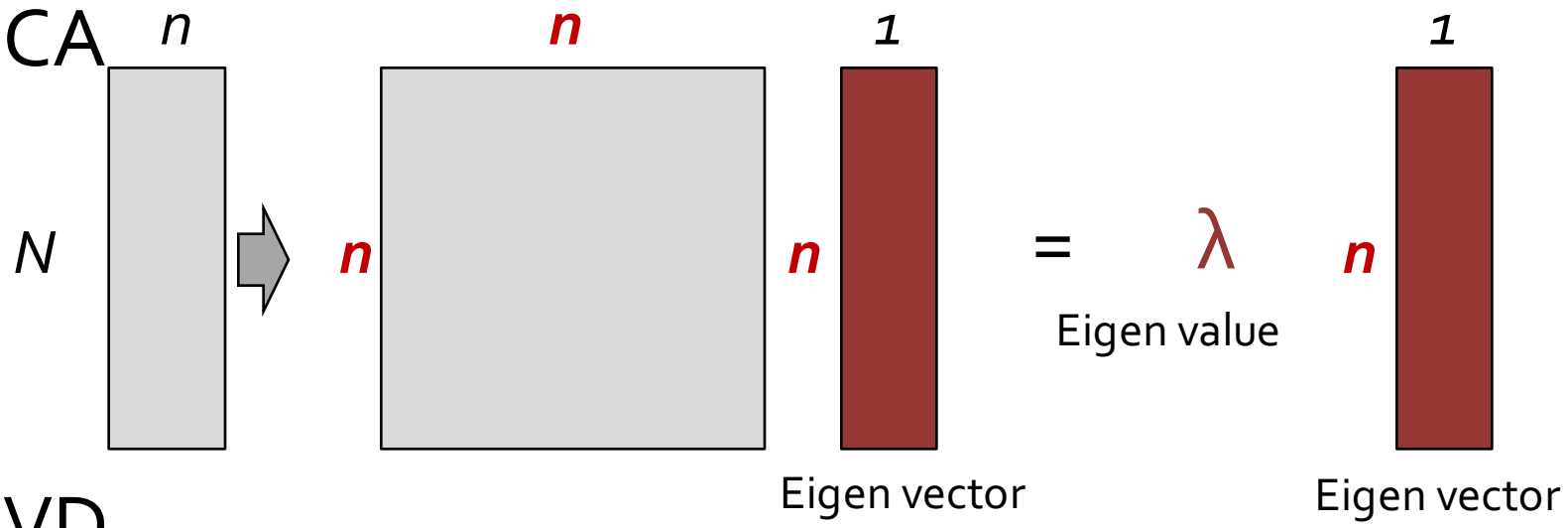
First we compute $\det(\mathbf{A} - \lambda\mathbf{I})$ via a cofactor expansion along the second column:

$$\begin{aligned} \begin{vmatrix} 7-\lambda & 0 & -3 \\ -9 & -2-\lambda & 3 \\ 18 & 0 & -8-\lambda \end{vmatrix} &= (-2-\lambda)(-1)^4 \begin{vmatrix} 7-\lambda & -3 \\ 18 & -8-\lambda \end{vmatrix} \\ &= -(2+\lambda)[(7-\lambda)(-8-\lambda)+54] \\ &= -(\lambda+2)(\lambda^2+\lambda-2) \\ &= -(\lambda+2)^2(\lambda-1). \end{aligned}$$

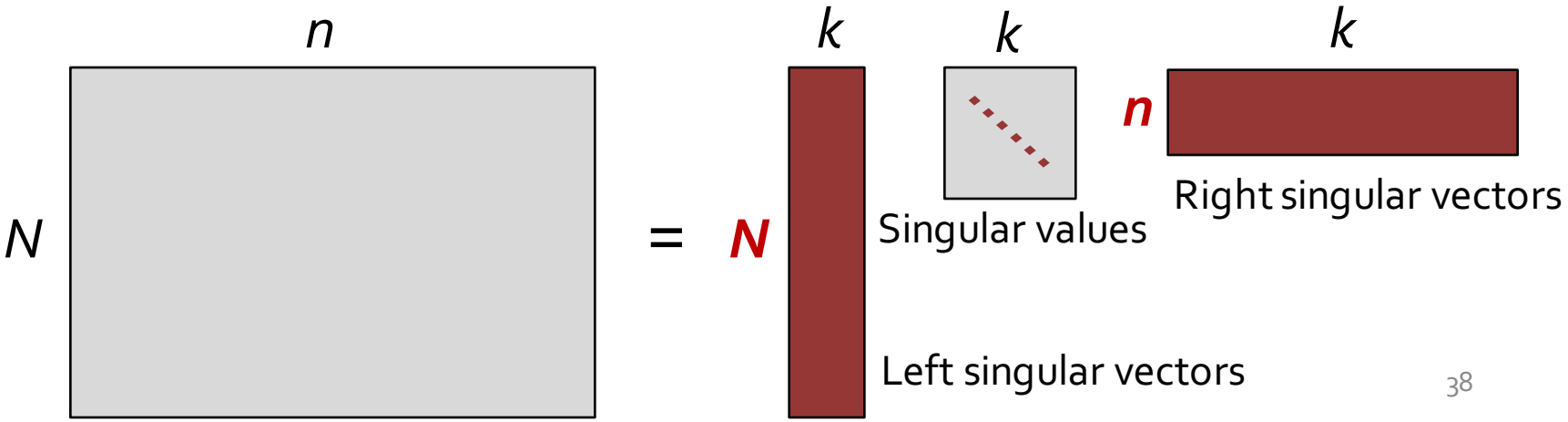
Thus $\mathbf{A}$ has two distinct eigenvalues, $\lambda_1 = -2$ and $\lambda_3 = 1$. (Note that we might say $\lambda_2 = -2$, since, as a root, -2 has multiplicity two. This is why we labelled the eigenvalue 1 as λ_3.)

Singular Value Decomposition (SVD)

- PCA



- SVD



Extra: Use SVD to Find Strange Behaviors

- Sell followers: "Become a Twitter Rockstar"

Jiang, M., Cui, P., Beutel, A., Faloutsos, C. and Yang, S., 2014, May. Inferring strange behavior from connectivity pattern in social networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 126-138). Springer, Cham.

A Strange Behavior

botnet

× 1,000

connect

customer

× 100

A Strange Behavior

botnet

× 1,000

connect

customer

× 100

#follower↑+1,000

A Strange Behavior

A Strange Behavior

botnet

customer

× 1,000

connect

× 100

More customers…

× 5,000

connect

× 100

A Strange Behavior

botnet

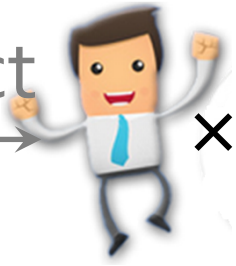
customer

× 1,000 connect

× 100

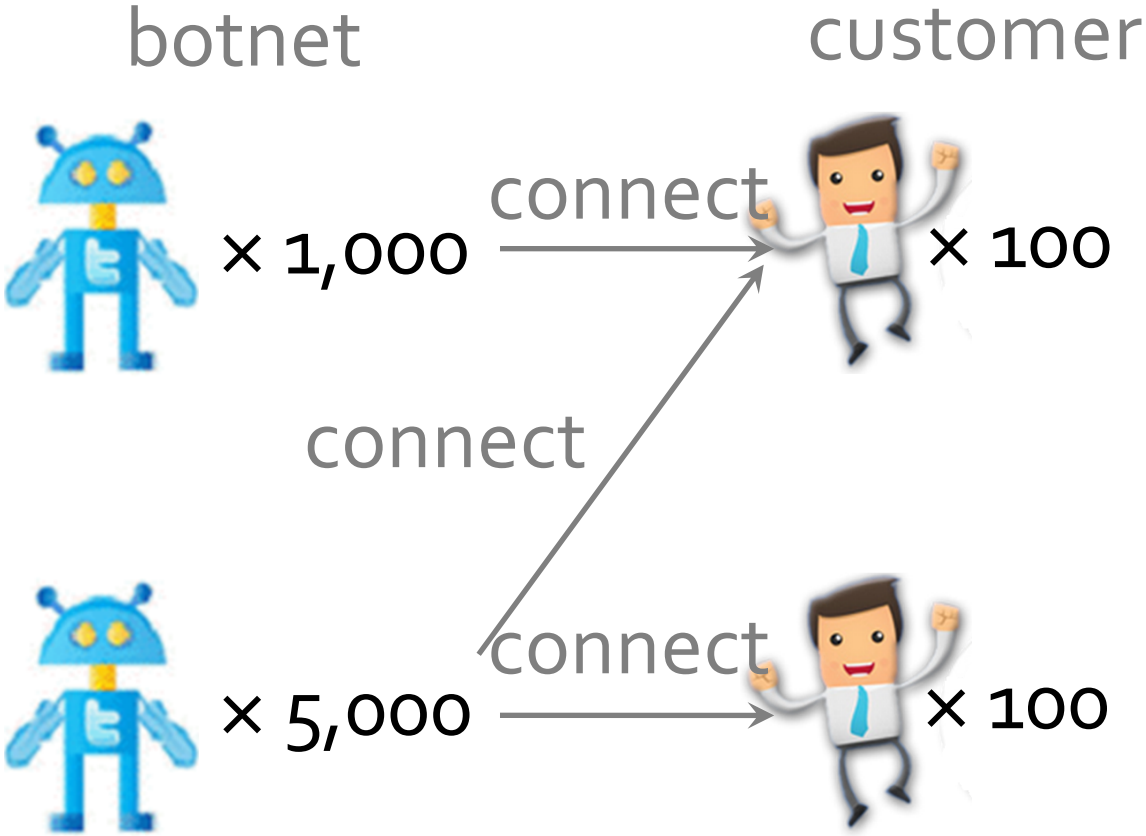
I want more followers…

× 5,000 connect

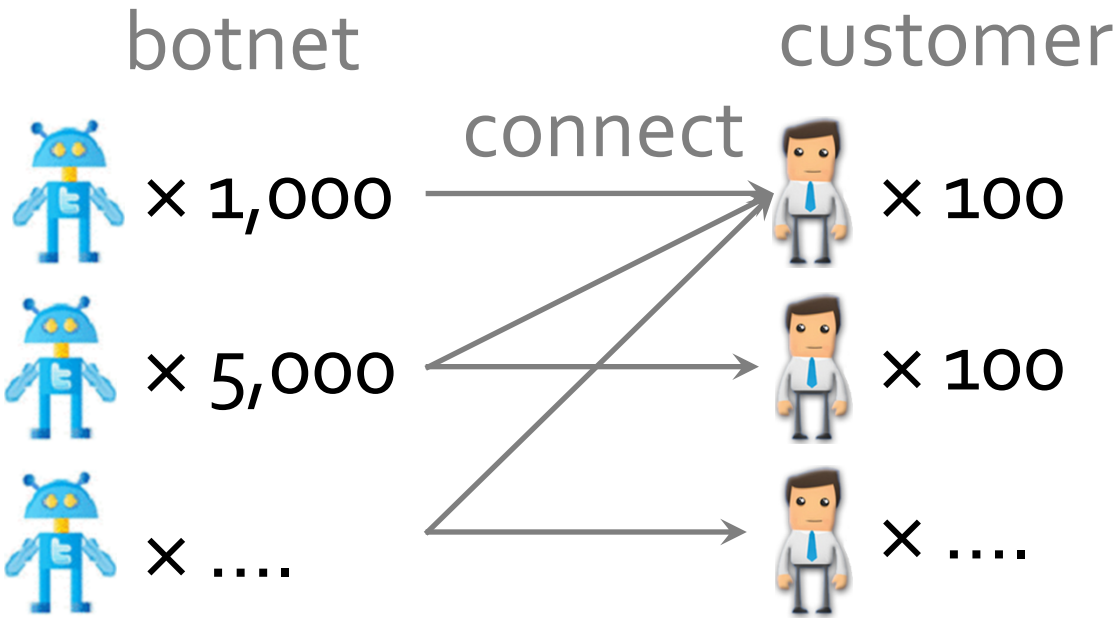


× 100

A Strange Behavior

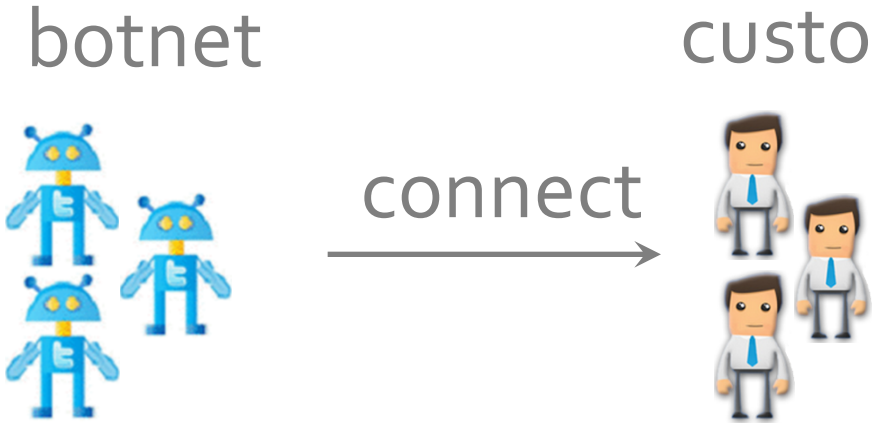


A Strange Behavior



More groups of customers
More groups of botnets
More companies....

A Strange Behavior



Detect dense biparitite cores!

How can we evade detection?

Some other activity!

A Strange Behavior

botnet

connect

customer

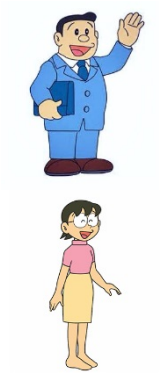
"Camouflage":
may connect
to popular idols
to look normal

A Strange Behavior

botnet

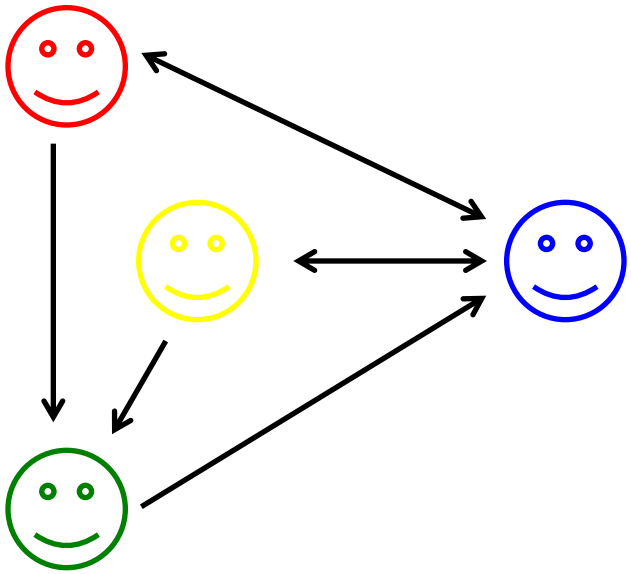
customer

connect

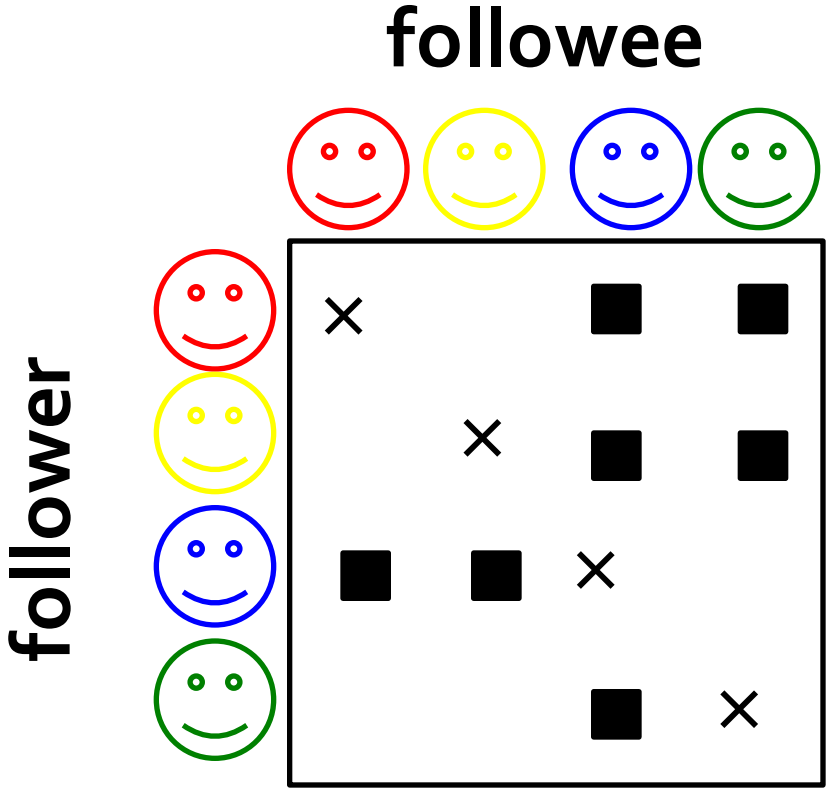


"Fame":
may have
a few honest
followers

Adjacency Matrix Reminder



Graph Structure



Adjacency Matrix

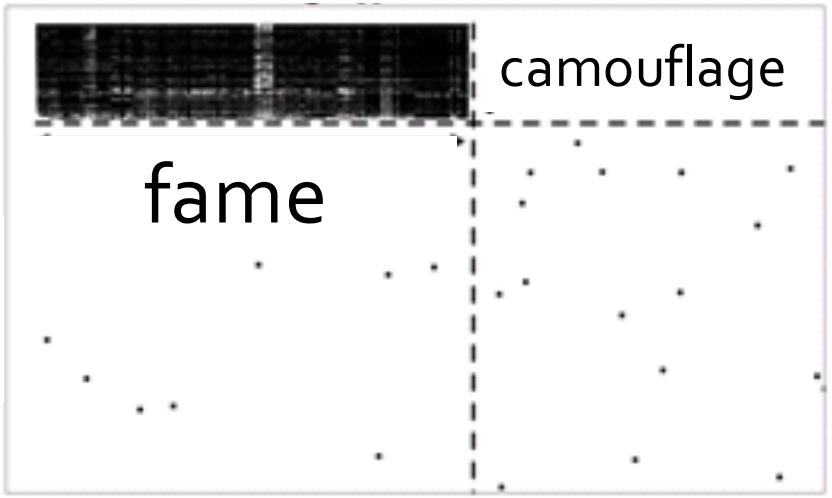
Strange ⟶ Lockstep Behavior

botnet

connect

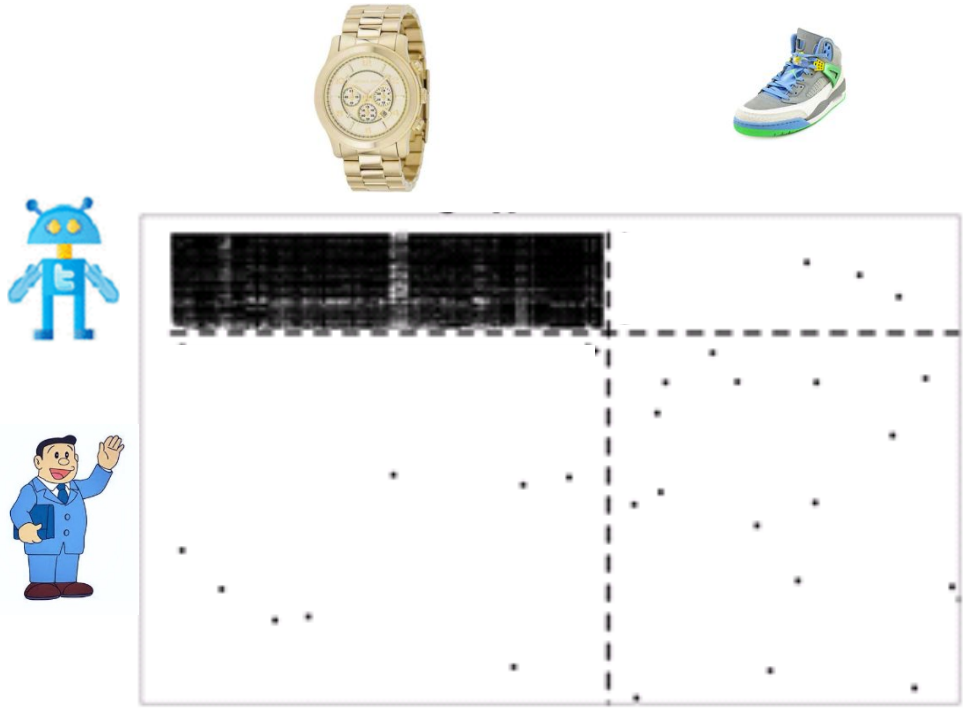
customer

- Groups
- Acting together
- Little other activity



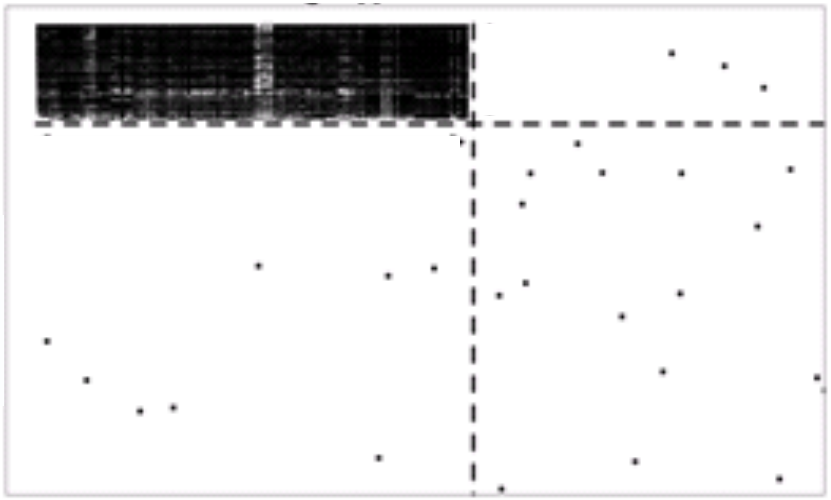
More Applications

- eBay reviews



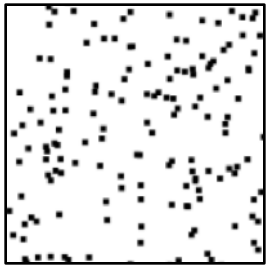
More Applications

- Facebook “Likes”

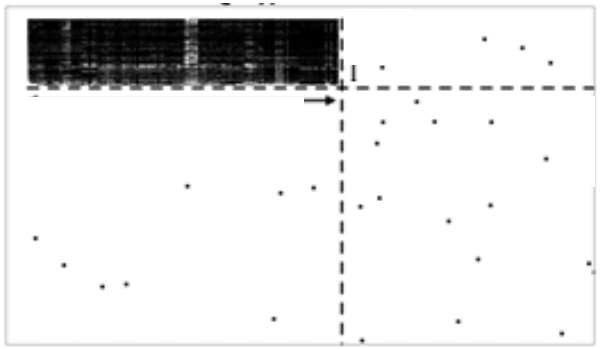


Problem Definition

- Given adjacency matrix

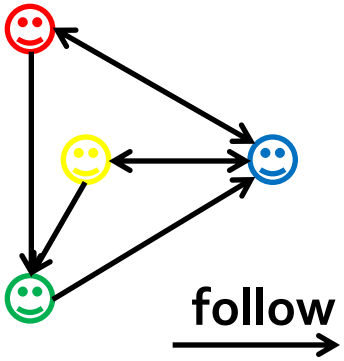


reordering

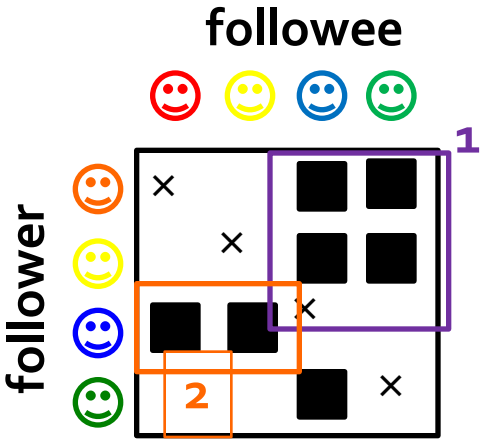


- Find Strange = "Lockstep" Behavior

SVD Reminder

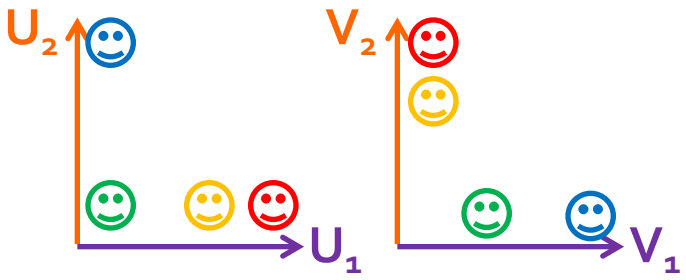


Graph Structure



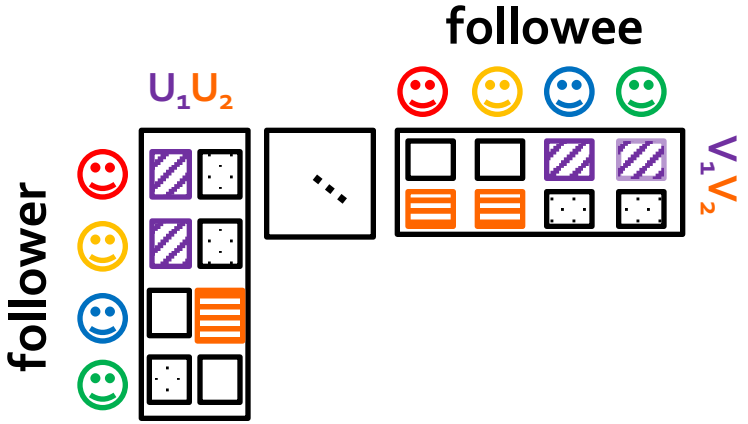
Adjacency Matrix

Pairs of singular vectors:



"Spectral Subspace Plot"

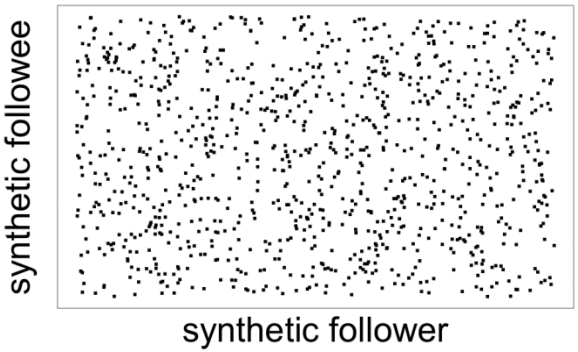
SVD: $A=USV^T$



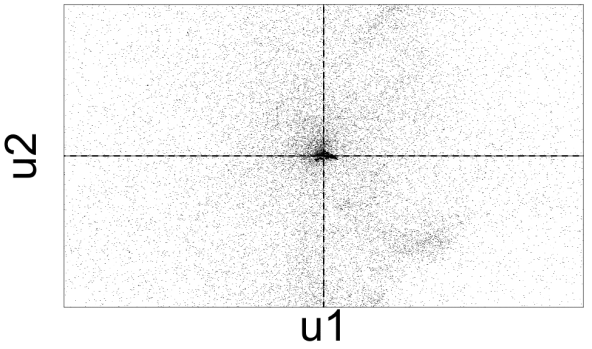
Lockstep and *Spectral Subspace Plot*

- Case #0: No lockstep behavior in random power law graph of 1M nodes, 3M edges
- Random ⟷ "Scatter"

Adjacency Matrix



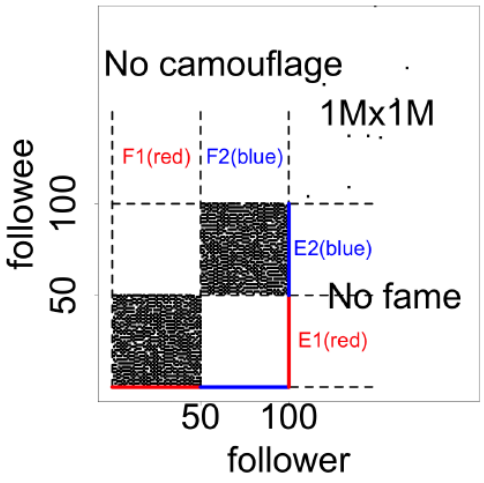
Spectral Subspace Plot



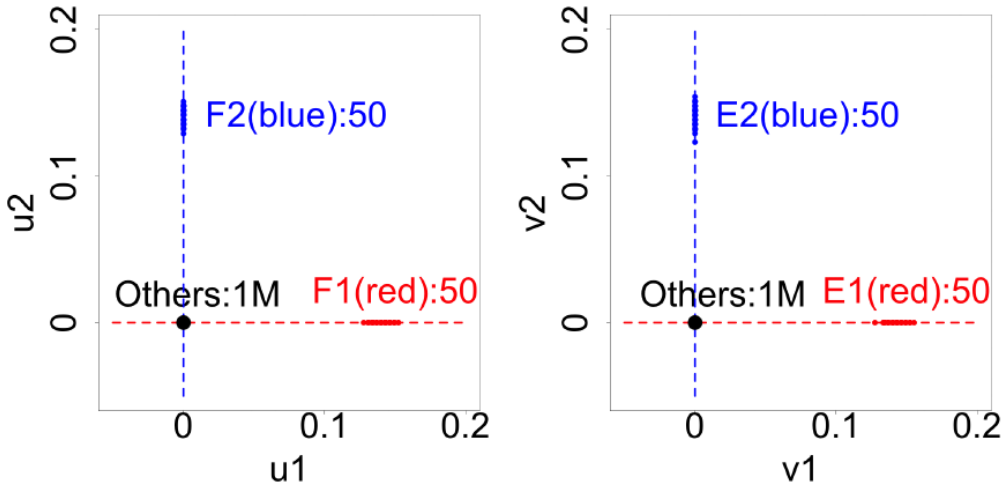
Lockstep and *Spectral Subspace Plot*

- Case #1: non-overlapping lockstep
- "Blocks" ⟷ "Rays"

Adjacency Matrix



Spectral Subspace Plot

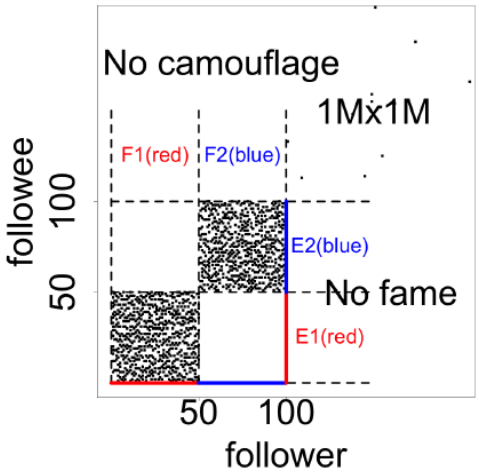


Rule 1 (short "rays"): two blocks, high density (90%), no "camouflage", no "fame"

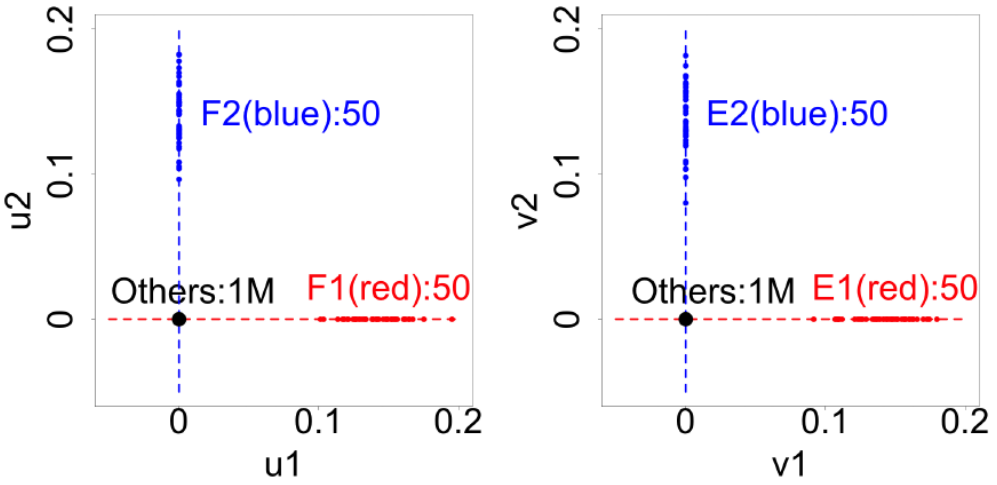
Lockstep and *Spectral Subspace Plot*

- Case #2: non-overlapping lockstep
- "Blocks; low density" ⟷ Elongation

Adjacency Matrix



Spectral Subspace Plot

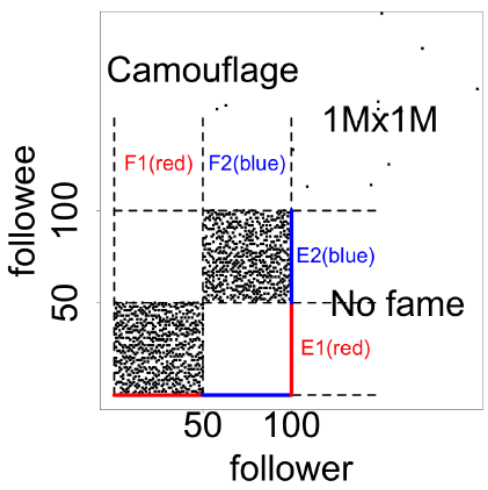


Rule 2 (long "rays"): two blocks, low density (50%), no "camouflage", no "fame"

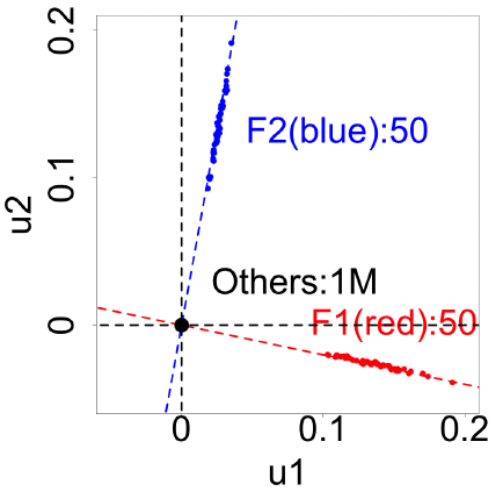
Lockstep and *Spectral Subspace Plot*

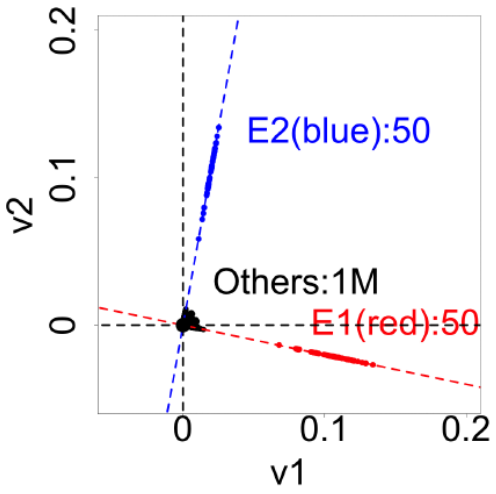
- Case #3: non-overlapping lockstep
- "**Camouflage**" (or "Fame") ⟷ Tilting "Rays"

Adjacency Matrix



Spectral Subspace Plot



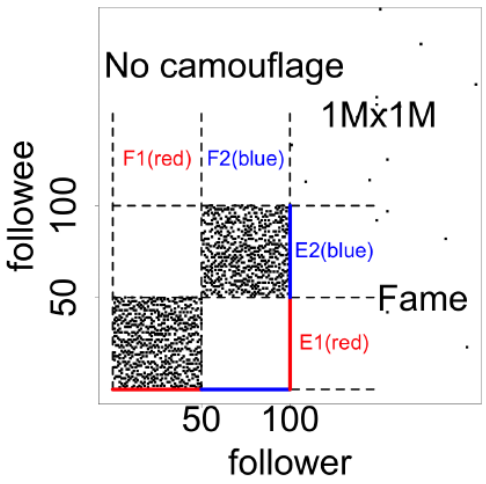


Rule 3 (tilting "rays"): two blocks, with "camouflage", no "fame"

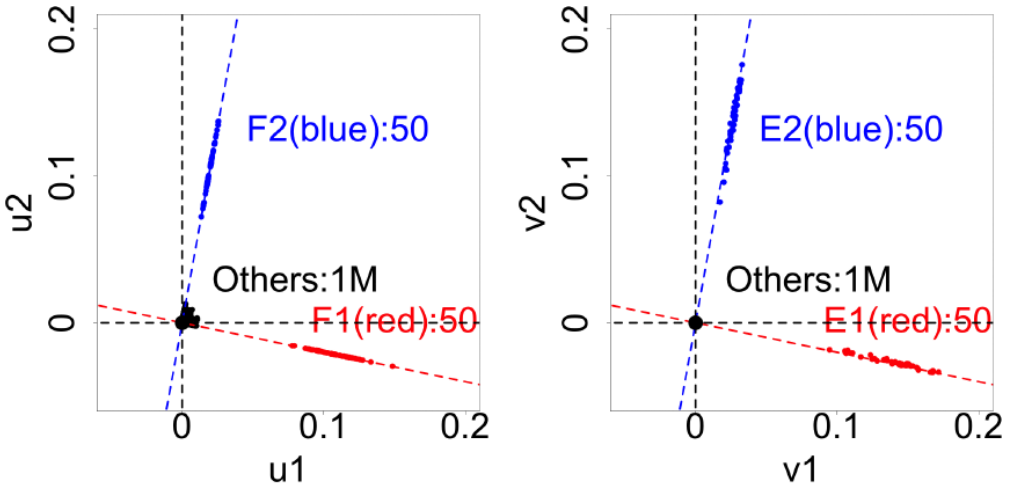
Lockstep and *Spectral Subspace Plot*

- Case #3: non-overlapping lockstep
- "Camouflage" (or "**Fame**")⟷Tilting "Rays"

Adjacency Matrix



Spectral Subspace Plot



Rule 3 (tilting "rays"): two blocks, no "camouflage", with "fame"

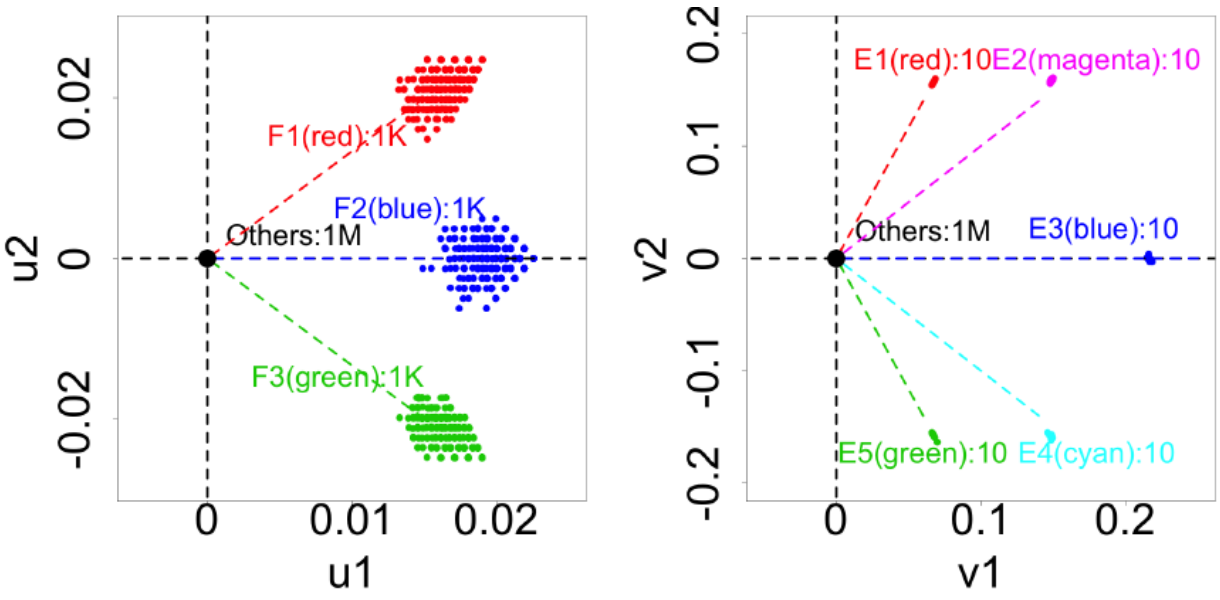
Lockstep and *Spectral Subspace Plot*

- Case #4: ? lockstep
- "?" ⟷ "Pearls"

Adjacency Matrix

?

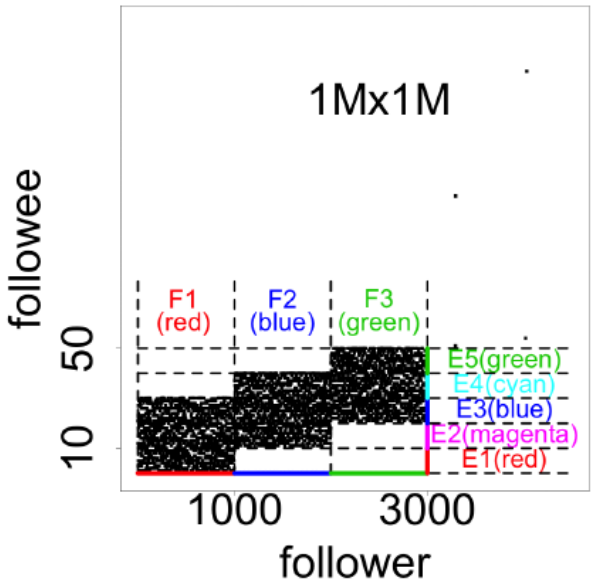
Spectral Subspace Plot



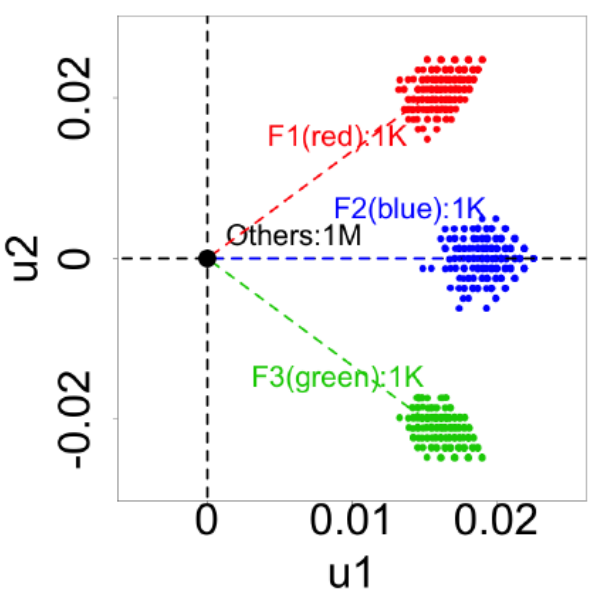
Lockstep and *Spectral Subspace Plot*

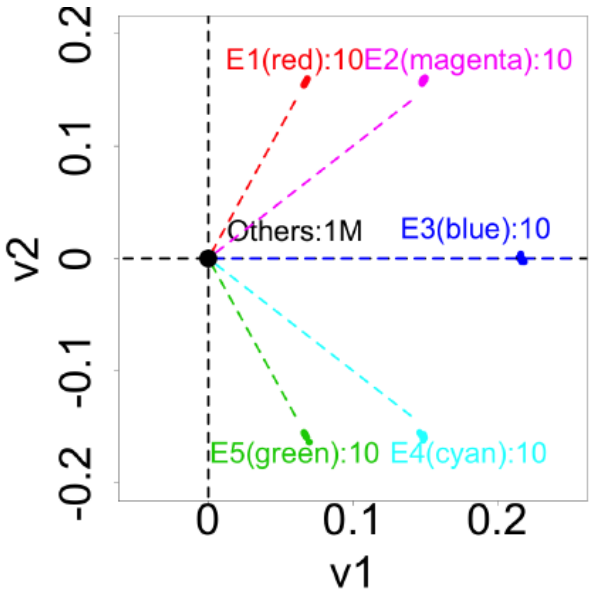
- Case #4: **overlapping** lockstep
- "**Staircase**" ⟷ "Pearls"

Adjacency Matrix



Spectral Subspace Plot





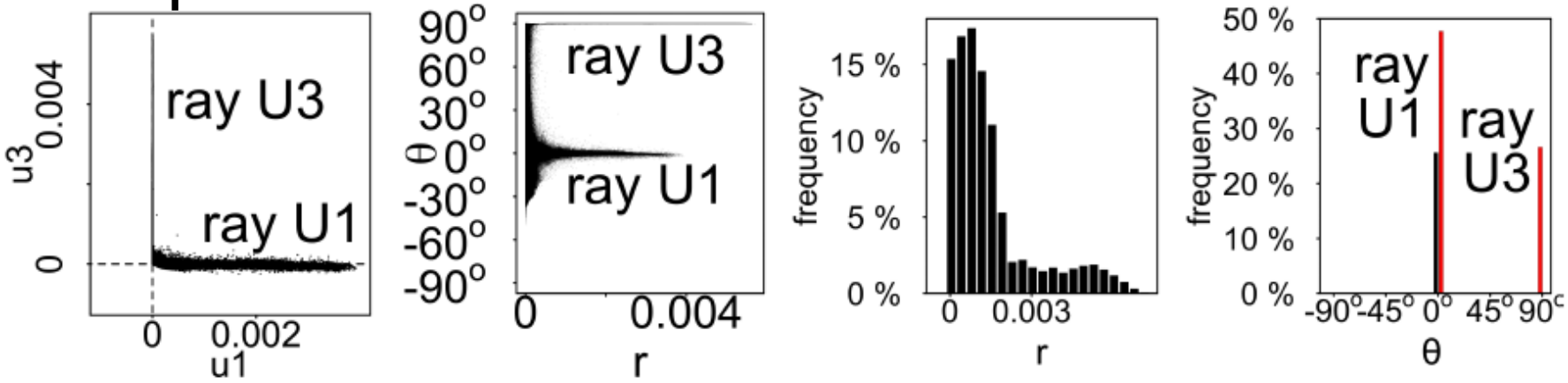
Rule 4 ("pearls"): a "staircase" of three partially overlapping blocks.

Algorithm

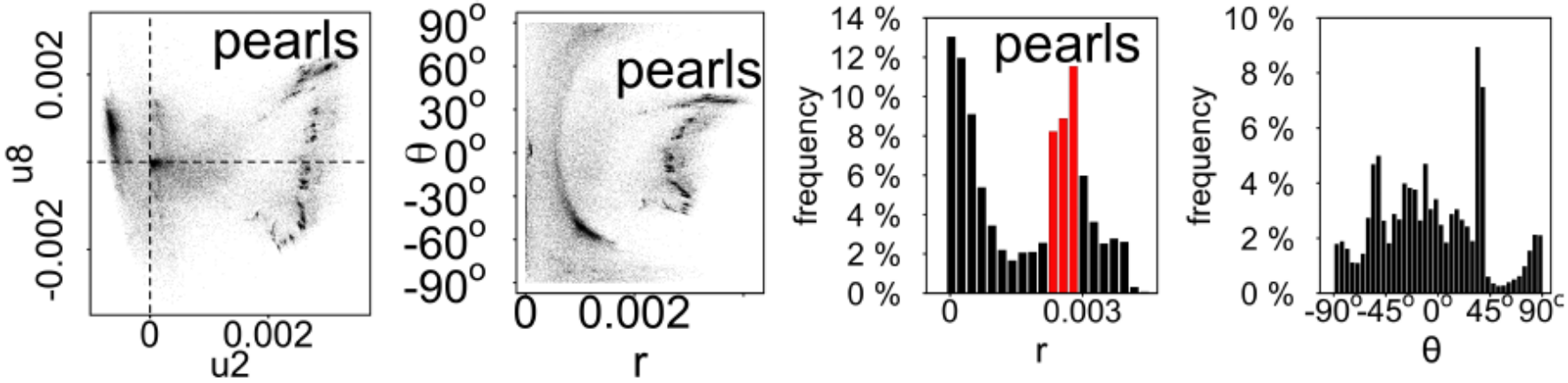
- Step 1: Seed selection
  - Spot “Rays” and “Pearls”
  - Catch seed followers
- Step 2: Belief Propagation
  - Blame followees with strange followers
  - Blame followers with strange followees

Automatically Spot "Rays" and "Pearls"

Spectral Subspace Plot | **Polar Coordinate Transform** | **Histograms**



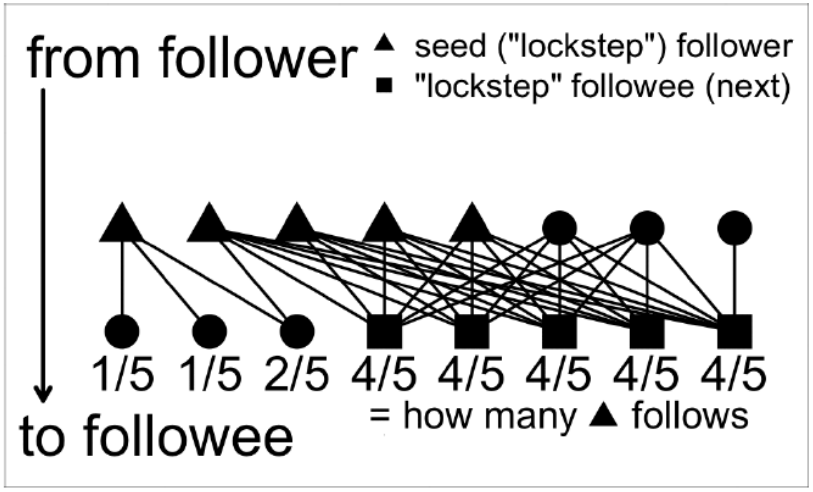
"rays" show two apparent spikes on θ frequency at 0^o and 90^o



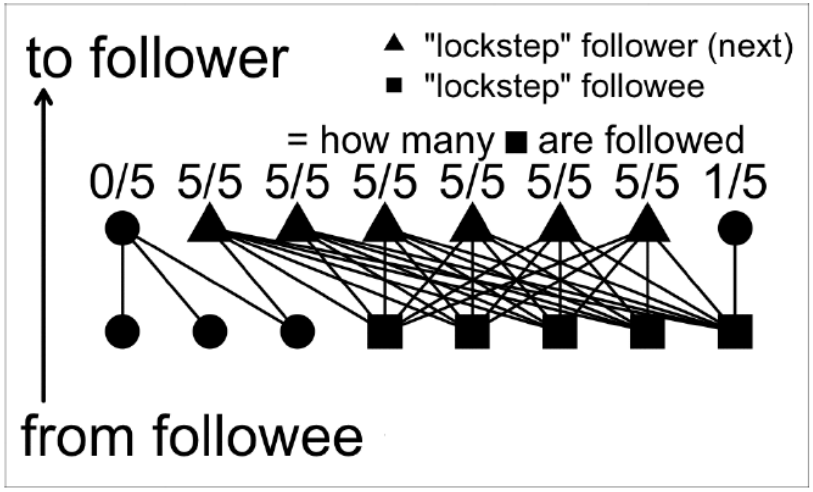
"pearls" show a spike on r frequency at a much-greater-than-zero value

BP-based Algorithm

- Blame followees with strange followers
- Blame followers with strange followees



(a) select "lockstep" followees: from (seed) followers to followees



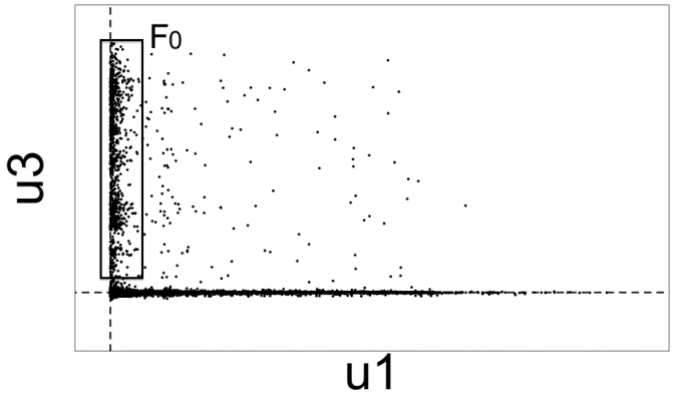
(b) select "lockstep" followers: from followees to followers

Dataset

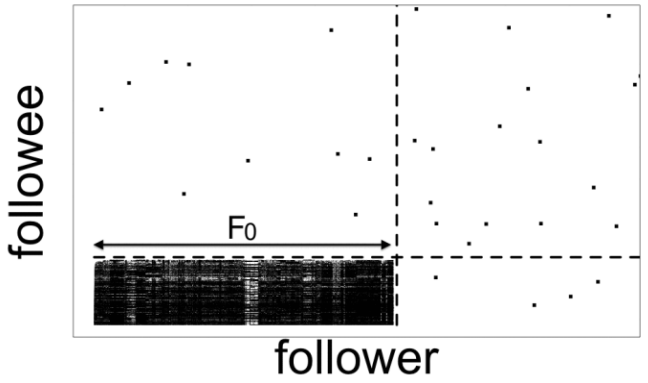
- Tencent Weibo 
- 117 million nodes (users)
- 3.33 billion directed edges

Real Data

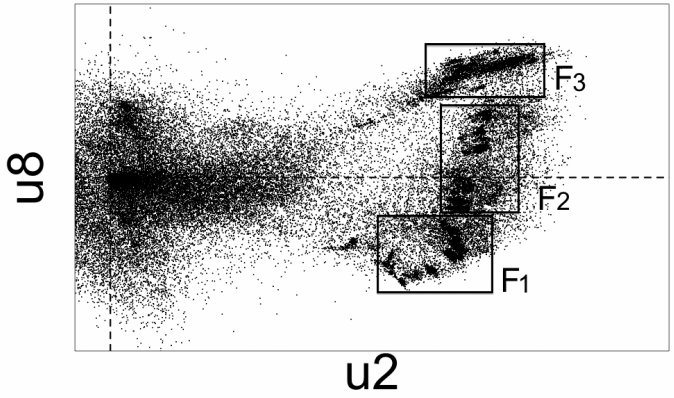
"Rays"



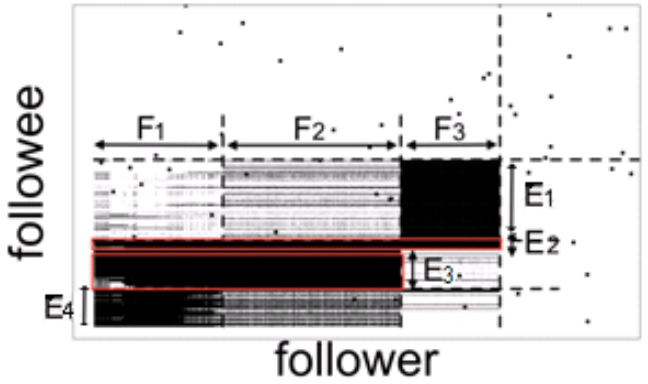
"Block"



"Pearls"

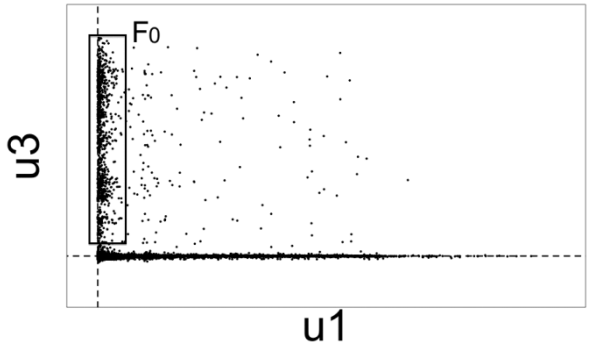


"Staircase"

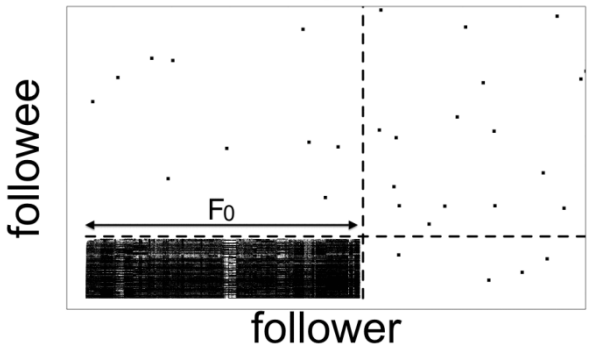


Real Data

"Rays"



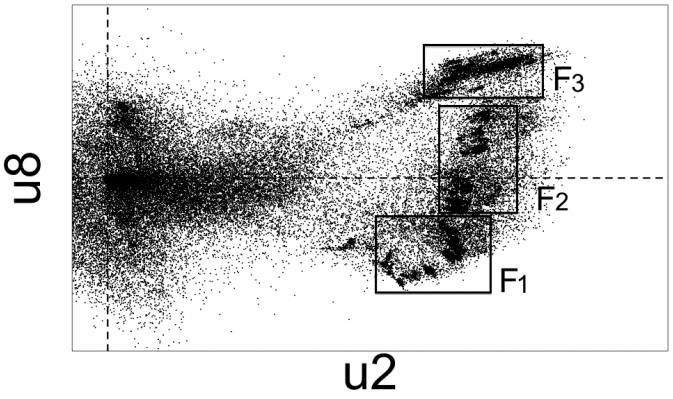
"Block"



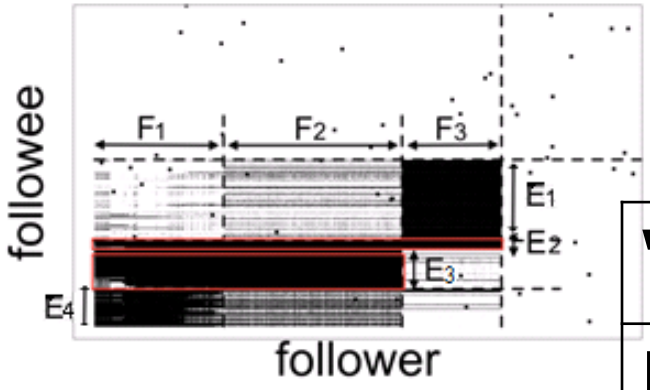
	"ray" F_0	"p
Num. seeds	100	
Size of block	$83,208 \times 30$	3, 1
Density	81.3%	
Camouflage	0.14%	
Fame	0.05%	
Out-degree	231 ± 109	
In-degree	2.0 ± 1.4	

Real Data

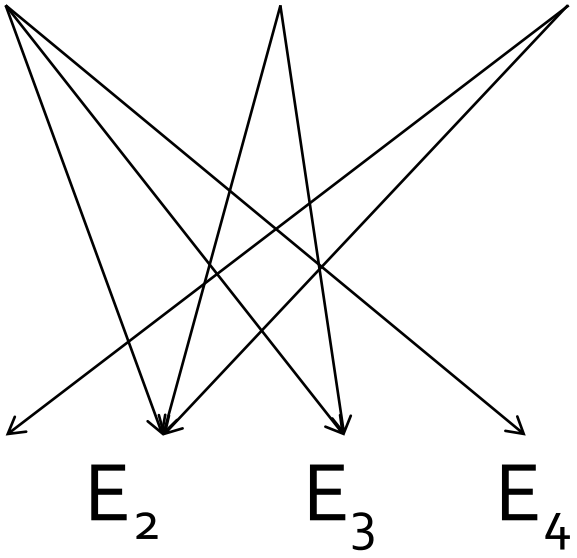
"Pearls"



"Staircase"



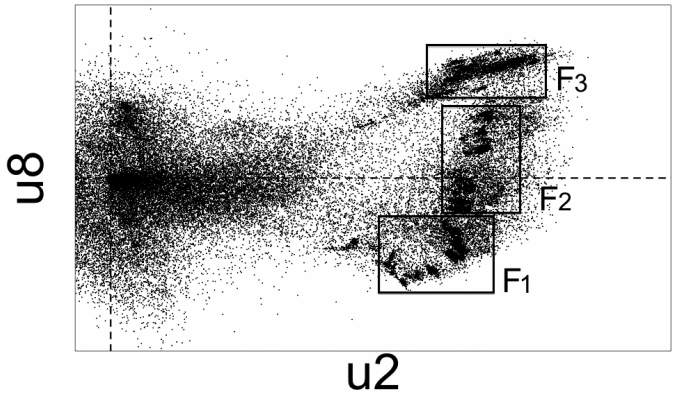
3,188 in F_1 7,210 in F_2 2,457 in F_3



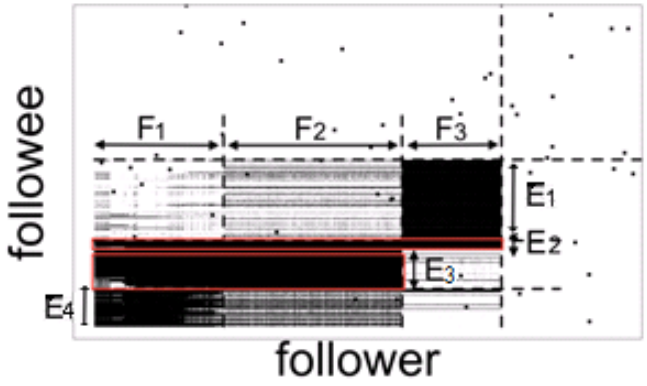
"F-E"	F_1-...	F_2-...	F_3-...
Density	91.3%	92.6%	89.1%

Real Data

"Pearls"



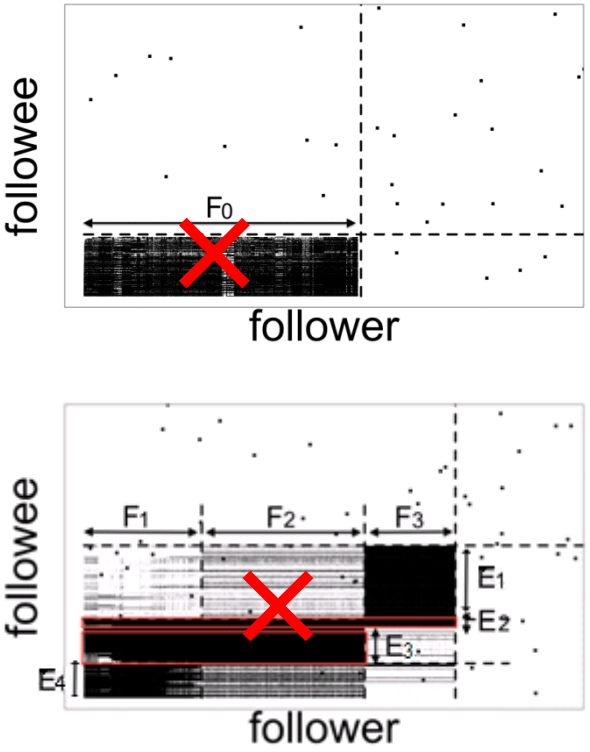
"Staircase"

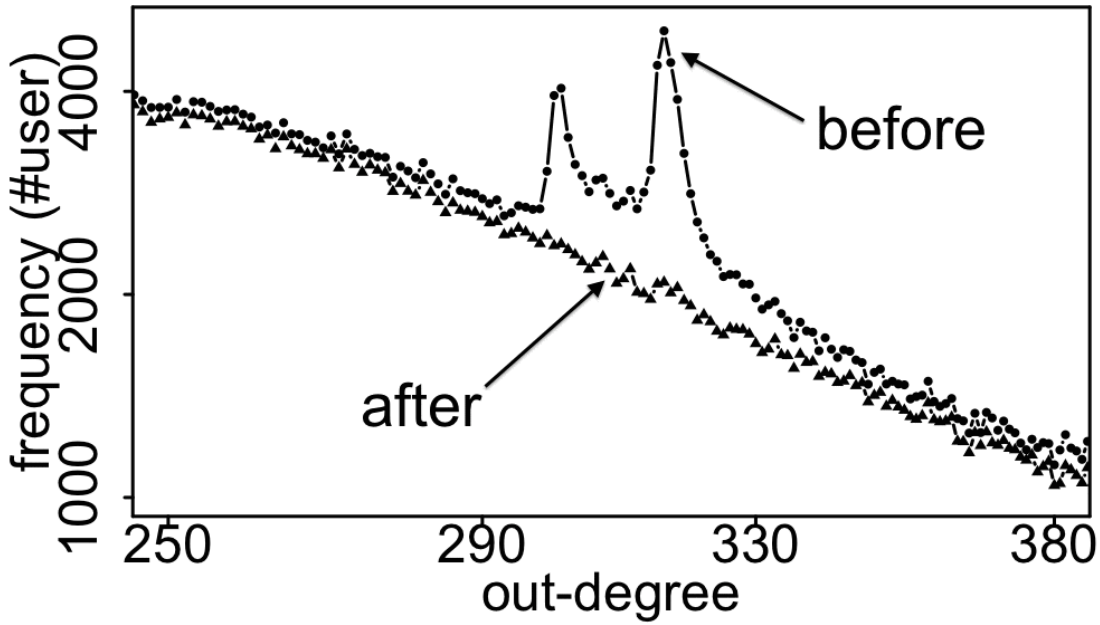


	"pearl" F_1	"pearl" F_2	"pearl" F_3	"pearl" Total
	1,239	107	990	—
	$3,188 \times 135$	$7,210 \times 79$	$2,457 \times 148$	$10,052 \times 270$
	91.3%	92.6%	89.1%	43.1%
	0.06%	0.10%	0.05%	0.07%
	1.93%	1.94%	1.72%	1.73%
	310±7	312±7	304±5	310±7
	[illegible]	[illegible]	[illegible]	[illegible]

Real Data

- Spikes on the out-degree distribution





Summary: Data Reduction

- Describe **numerosity reduction** (reducing #instances)
  - Parametric methods: Fit some model and estimate model parameters
    - Regression: Describe linear/non-linear regression models
  - Nonparametric methods
    - Histograms
    - Clustering
    - Sampling: Describe stratified sampling
- Describe **dimensionality reduction** (reducing #features)
  - Feature selection
    - Heuristic search
  - Feature extraction
    - Principal component analysis (PCA)
    - Singular Value Decomposition (SVD)

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995