Dynamic and Social Network Analysis

Lecture 8

Miray Kas
Bilkent University
Computer Engineering Department

Poll for the 5th homework

- **Option-1**: Convert attendance to 5th homework.
- **Option-2**: Regular take home homework (similar size to prior homeworks)
- **Option-3**: In-class informed quiz style (not pop-up, agreed time/date) (10-15 True/False Questions)

Let’s do a poll!

⇒ Poll results: Option-1, Convert attendance to Assignment-5. For those that are interested, a homework (option-2) will be released.

Midterm Exam

- Midterms are graded.
  - I will try to upload grades to Moodle. If I cannot, I will share the grades via a spreadsheet
  - We will arrange a time with your TA to show the papers
  - Need to give make up exam to a student, the time to show the papers will be arranged after that.

Next Project Milestone

- The deadline was this Wednesday, it is now extended to December 3rd, Saturday.

Dynamic Networks

Dynamic Network Analysis

- This is an emerging field of analysis
- Different frameworks exist to handle dynamic networks
  - Dynamic Metanetworks is one such framework.
- ORA depends on and implements Dynamic Metanetworks framework
- This concept is explained in detail here

Metanetworks

- **Meta Network:** A multi-mode, multilink, multi-level network
  - Unit of analysis for Dynamic Network Analysis (DNA)
- **Dynamic meta-network** is a structured collection of meta-networks.
  - Provides a way of grouping meta-networks together to record network evolution or change.
- **Simulations:**
  - Agent-based modeling and other forms of simulations are often used to explore
    - How networks evolve and adapt
    - Impact of induced changes (interventions) on those networks.

Dynamic Meta-Network Components

- Node → NodeSet → Network → Metanetwork → Dynamic Metanetwork
- **Metanetwork**
  - A meta-network incorporates both nodesets and networks into a unit.
  - Unit of analysis for dynamic network analysis
  - Each metanetwork may be given an optional timestamp
- **Dynamic Metanetwork**
  - container for meta-networks
  - Uses the meta-network time value to indicate ordering

Keyframe (Snapshot) vs. Delta

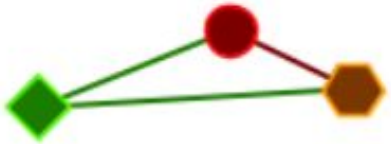
- **Keyframe**
  - A starting point, or a snapshot of what the meta-network looks like at any given time.
  - Doesn't care what came before or will occur after.
  - Keyframes contain every node and network for each time slice.
- **Delta**
  - A set of instructions on how to change the metanetwork.
  - It reviews what came before it, applies a set of changes, and displays the revised metanetwork.
  - Advantage: Requires less information and space than full snapshot
  - Only stores information about individual changes between time slices

Keyframe (Snapshot) vs. Delta

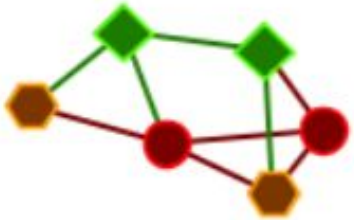
Keyframe Representation

Initial State : Nothing

KeyFrame : 3 nodes, 3 links



KeyFrame : 6 nodes, 9 links



Delta Representation

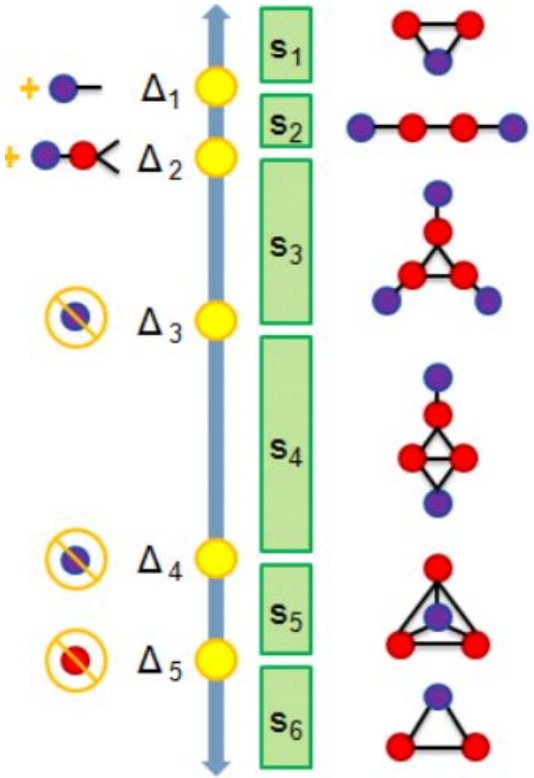
Initial State : Nothing

Delta : +3 nodes, +3 links

Delta : +3 nodes, +6 links

Img src

Conceptual Dynamic Metanetwork Example



- S1 : 2 agents share 1 location
- S2 : They move apart to separate locations
- S3 : A 3rd connected agent is detected at its own location
- S4 : 2 of the agents meet
- S5 : The 3rd joins at same locale
- S6 : One agent disappears

6 meta networks at 6 consecutive time periods

Img src

Time Notation

- Time notation dictates ordering, so it is important to be consistent
- Does not have to be timestamp, could be freeform strings but order must be correct

Time Format	Time Type	Example
yyyy-MM-dd'T'HH:mm:ssX	Date	2001-07-04T12:08:56-0700
yyyy-MM-dd HH:mm:ssX	Date	2001-07-04 12:08:56-0700
yyyy-MM-dd HH:mm:ss	Date	2001-07-04 12:08:56
yyyy-MM-dd	Date	2001-07-04
yyyy-MM	Date	2001-07
MM/dd/yyyy	Date	07/04/2001
EEE MMM dd HH:mm:ss ZZZZ yyyy	Date	Sat Jul 04 12:08:56 -0700 2001
Time period string	Period	Mesozoic

Img src

Aggregation for Metanetworks in DNA

- Allows creation of a metanetworks (and dynamic meta-networks) where a range of time are merged into a single meta-network.
  - Aggregate email messages by day
  - Seasonal disease mortality.
  - Publications by year.

Aggregations are commonly done via standard time units

- Year
- Month
- Week - Sunday Start
- Week - Monday Start
- Day
- Hour
- Minute
- Second

Comparison Techniques

Comparison Techniques

- Networks differ from snapshot to snapshot.
- How do you compare the two?
- Common comparison techniques
  - Distances
  - Similarities
  - Regression
  - Pearson Coefficient
  - QAP/MRQAP methods

Comparison Techniques: Distances

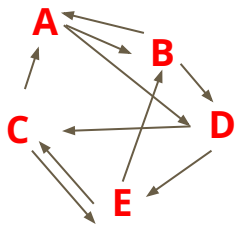
- Assume each network is a string.
- Also relevant for Machine Learning and clustering
  - Algorithms like k-means have distance functions at their heart.
- Most famous techniques:
  - Hamming
  - Euclidean
  - Manhattan
  - Chebyshev

Hamming Distance

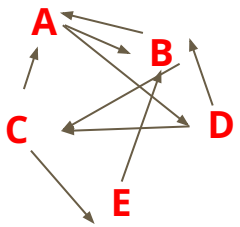
- Binary computation
- Calculate the distance (or difference) between the two strings
- Multiple ways of saying the same thing:
  - How many bits do you need to flip in adjacency matrix A to make it look like adjacency matrix B?
  - The number of bits which differ in two binary strings
  - Union (A_i, B_i) – A_i

Hamming Visual Example

	A	B	C	D	E
A	0	1	0	1	0
B	1	0	0	1	0
C	1	0	0	0	1
D	0	0	1	0	1
E	0	1	1	0	0



	A	B	C	D	E
A	0	1	0	1	0
B	1	0	1	0	0
C	1	0	0	0	1
D	0	1	1	0	0
E	0	1	0	0	0



0 1 0 1 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 1 1 0 0

0 1 0 1 0 1 0 1 0 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0

Calculation

5/(5*(5-1)) = 5/20⇒ 25%

PROCESS: Picture ⇒ Matrix ⇒ String ⇒ Calculation

Hamming Distance, Difference, Similarity

- **Maximum Hamming Distance** = N (N-1)
- **Hamming Difference**
  - Since different networks have different sizes, it is more useful as a percentage and more meaningful across different networks.

$$HammingDifference = 100 * \frac{(MaxPossibleDistance - Hamming)}{MaxPossibleDistance}$$

- **Hamming Similarity**

$$HammingSimilarity = 1 - HammingDifference$$

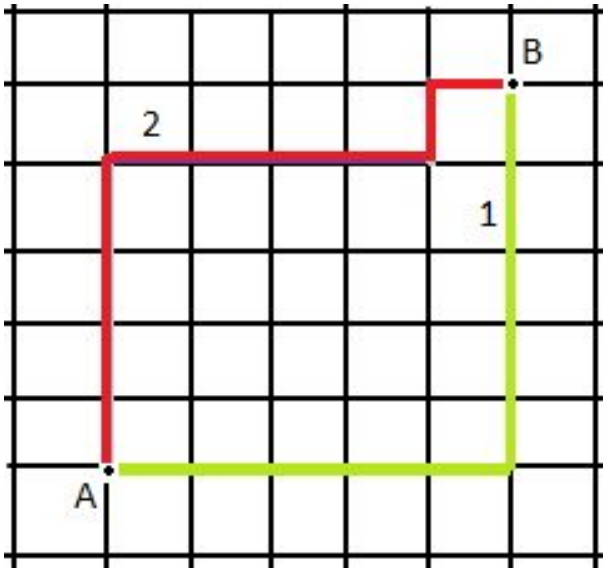
Euclidean Distance

- Can be used with valued or binary data
- Physical distance
- How to compute:
  - The strength of node-A's tie to node-C is subtracted from the strength of node-B's tie to node-C, and the difference is squared.
  - This is repeated across all the other nodes (D, E, F, etc.), and summed.
  - The square root of the sum is then taken.

$$EucledianDistance(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

Manhattan Distance

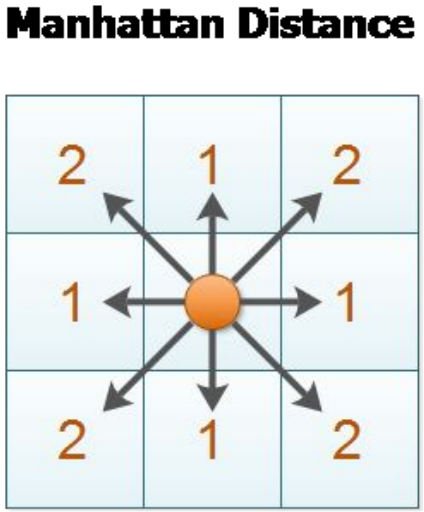
- Sum of the absolute differences between the two vectors.
- Follows only axis aligned directions
- Block distance between the two vectors



- Better for agent models built on grid space
- Also good for geospatial networks, they naturally come with latitude/longitude based grids

Manhattan Distance

- This distance is simply the sum of the absolute difference between the actor's ties to each alter, summed across the alters.

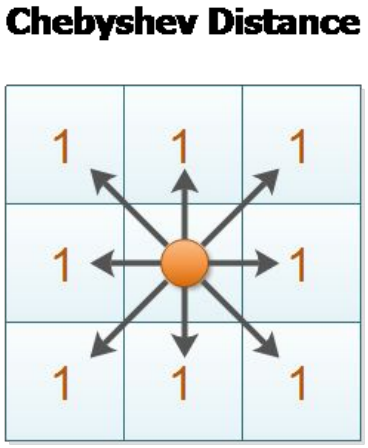


$$|x_1 - x_2| + |y_1 - y_2|$$

Img src

Chebyshev Distance

- Uses only the most significant dimension
- Metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension.

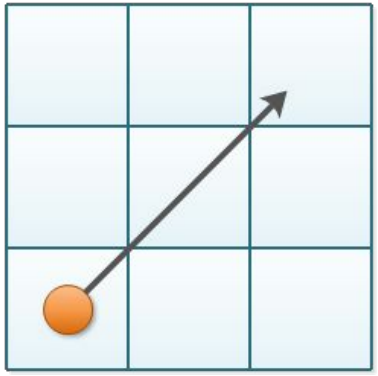


$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

Img src

Euclidean / Manhattan / Chebyshev Distances

Euclidean Distance

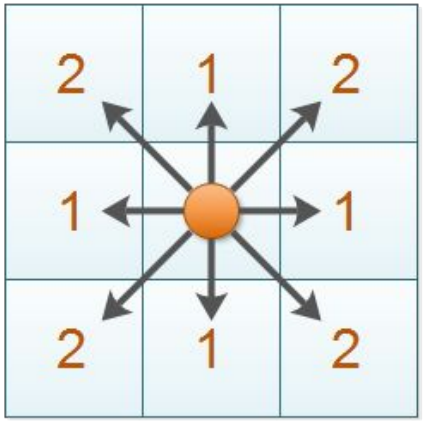


$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

(Straight Line)

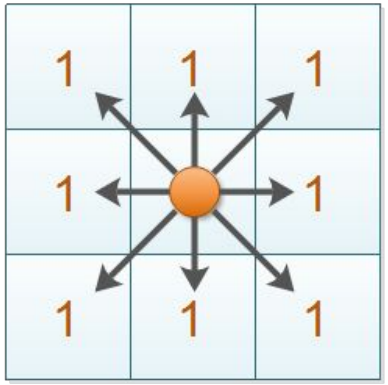
(City Block)

Manhattan Distance



$$|x_1 - x_2| + |y_1 - y_2|$$

Chebyshev Distance



$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

(Chessboard)

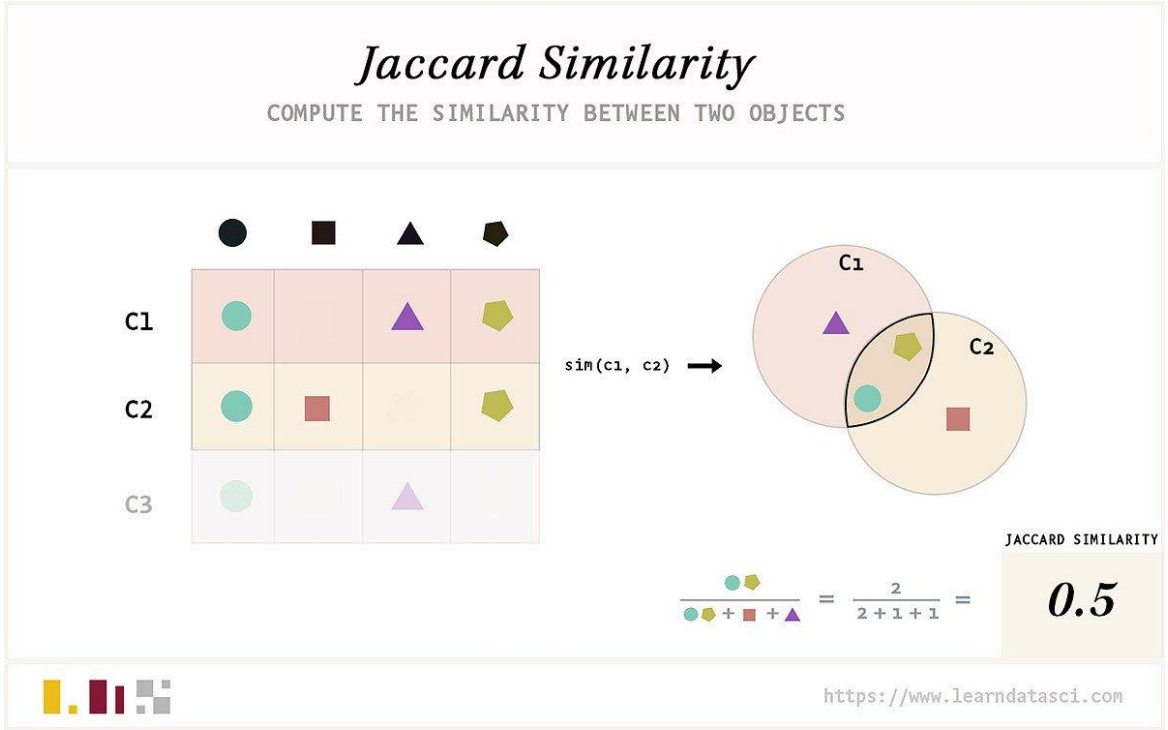
Img src

Comparison Techniques: Similarities

- Hamming-based similarity (discussed)
- Jaccard Similarity
- Sorensen Similarity
- Pearson’s Correlation coefficient

Jaccard Similarity (Intersection over Union)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



Sorensen Similarity

- The Sørensen similarity **equals twice the number of elements common to both sets divided by the sum of the number of elements in each set**.

$$S(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

- It resembles Jaccard similarity, but gives more weight to the common elements

Network Comparison Techniques

- **Regression Analysis**
- **QAP** - **Q**uadratic **A**ssignment **P**rocedure
  - Pearson’s coefficient
- **MRQAP** - **M**ulti **R**egression **Q**uadratic **A**ssignment **P**rocedure
  - Regression on Networks

- Useful for comparing multiple networks
  - **QAP:** Good for comparing two networks.
  - **MRQAP:** Good when there are more than two networks.

What can you answer by comparing multiple networks?

- Do marriage ties correlate with business ties in the Medici family network?
- Are friendship relations correlated with work relations?
- Do the trends in 1990s correlate with 2000s?

Regression Analysis for Networks

- Is one network a function of another network?

- Is the perceived friendship network a function of the actual contact network?

Src

Regression Analysis

- Regression assumes that one variable (dependent) is a function of another variable (independent)
  - A set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables
- Takes data on variables and determines the values of the coefficients;
- Assesses how confident we can be in those estimates
- Determines the coefficients by finding the best fitting line through the data
- The function is then found by estimating the conditional expectation

Src

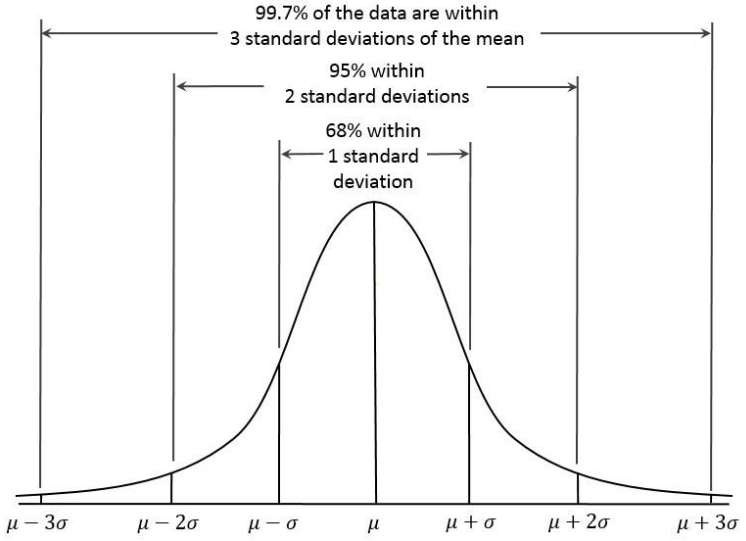
Mini Recap: Confidence Level

- The confidence that the researcher has that the selected sample is one that estimates the population parameter to within an acceptable range
- Usually expressed as the probability that a parameter lies within some range of the sample statistic
  - Range is called the confidence interval and is usually expressed in terms of the standard error
- 95% confidence is typical in research
  - The more confident we want to be, the more data is needed

Src

Mini Recap: Standard Error

- Measures the accuracy of a sample – Expresses how close the sample statistic is to the population parameter
- **If the standard error is small,** then the sample estimates based on that sample size will tend to be similar and will be close to the population parameter
- **If the standard error is large,** then the sample estimates will tend to be different and many will not be close to the population parameter



Text src

Mini Recap: Coefficient of Determination

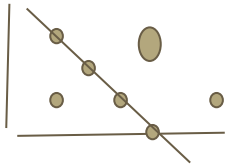
- **R^2** measures the percentage of total variation in the dependent variable (Y) that is explained by the regression equation
- Ranges from **0 to 1**
- **High** R^2 indicates Y and X are **highly correlated**

Src

Linear Regression Analysis

- Simple linear regression relates dependent variable Y to one independent (or explanatory) variable

$$X - Y = \mathbf{a} + \mathbf{b}X$$



Intercept parameter (a) gives the value of Y where regression line crosses Y-axis (value of Y when X is zero)

Slope parameter (b) gives the change in Y associated with a one-unit change in X

- Parameter estimates are obtained by choosing values that minimize the sum of squared residuals
- The residual is the difference between the actual and fitted values of Y – Called Ordinary Least Squares (OLS)

Src

Dependency Issue in Networks

- Unit of analysis is a dyad (e.g. a pair)

- The problem is that **the observations are not independent on each other**.
  - If A is related to B, and B is related to C, it may be relatively likely that A is related to C.

- The independent variables would be either attributes of each of one or both members of the pairs, or of similarities and / or matches between the pairs.

Problems with Statistics on Networks

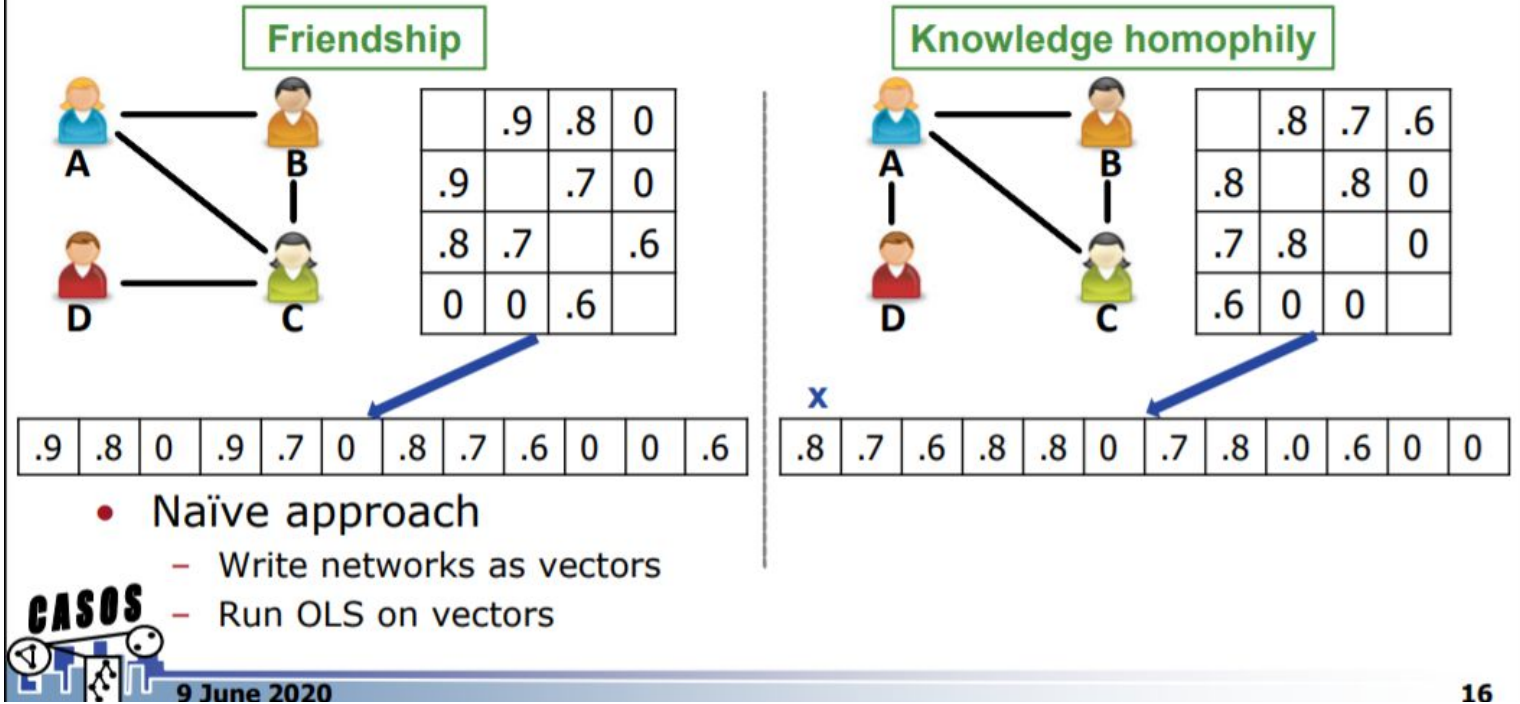
- The basic assumptions of standard statistics are violated
  - Estimation procedures designed for independent

- In networks, there are row / column dependencies
  - Each entry is a dyad and dyads are not independent
  - Observations are correlated

- Observations will calculate incorrect standard errors
  - The fact that there are repeating observations means that the errors are correlated with each other.
  - Observations in individual rows or in individual columns tend to be highly correlated. This inflates or deflates standard errors.

Problems with Statistics on Networks

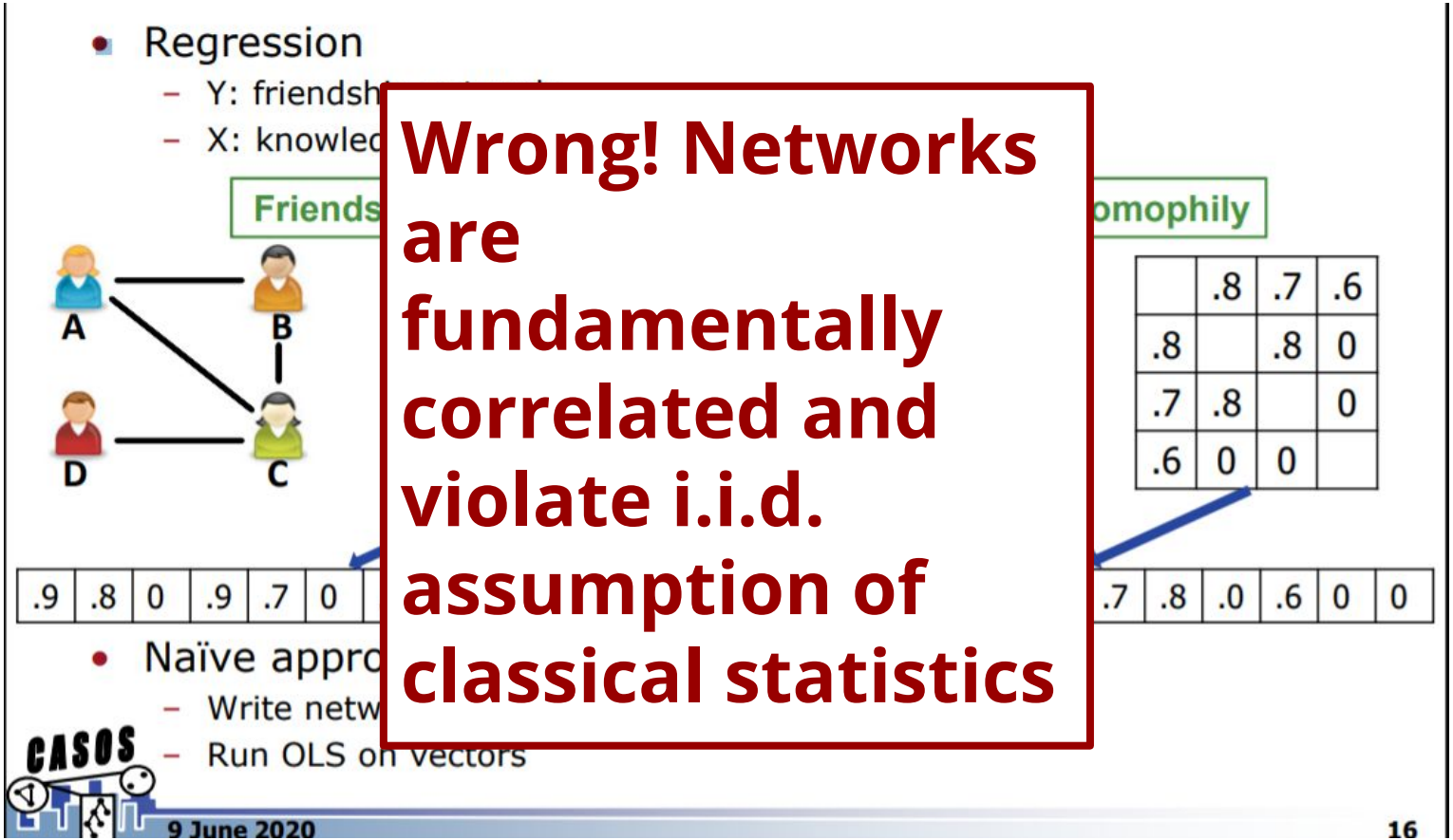
- Regression
  - Y: friendship network
  - X: knowledge homophily network



- Naïve approach
  - Write networks as vectors
  - Run OLS on vectors

src

Problems with Statistics on Networks



Wrong! Networks are fundamentally correlated and violate i.i.d. assumption of classical statistics

src

QAP - Quadratic Assignment Procedure

- A permutation test that controls for this independence problem

- Scramble dependent variable data through several permutations
  - By scrambling the data repeated it results in several random datasets with the dependent variable
  - Those datasets form an empirical sampling distribution
  - Then, multiple analyses can be performed.

- Standard errors are estimated by using permutations of the data set.

Think about tossing a coin!

- Repetition helps!

- If you toss it enough many times, you get closer to the expected value (0.5 Head, 0.5 Tail)

QAP - Quadratic Assignment Procedure

- This test is performed by:
  - Repeatedly (randomly) relabeling the input graphs
  - Recalculating the test statistic
  - Evaluating the fraction of draws >= OR <= the observed value

- Preserve row-column dependencies
  - For a single node, the row and column remain the same, and are permuted in the same way, so that the rows and columns for a single node are not separated.

QAP - Quadratic Assignment Procedure

1. Calculate the Pearson's correlation coefficient (r) between the two original matrices, treating the computed coefficient as the observed coefficient.
2. Permute the dependent matrix Y by rearranging both its rows and columns.
3. The permuted Y matrix is correlated with the original independent X matrix, producing a new Pearson's correlation coefficient (r) between the two matrices.
4. Steps 2 and 3 are repeated at least 1000 times.
5. The observed correlation coefficient from the first step is compared with the distribution of the coefficients generated from step 4 to determine the proportion among the coefficients from the permuted matrices that are equivalent or higher than the observed coefficient.

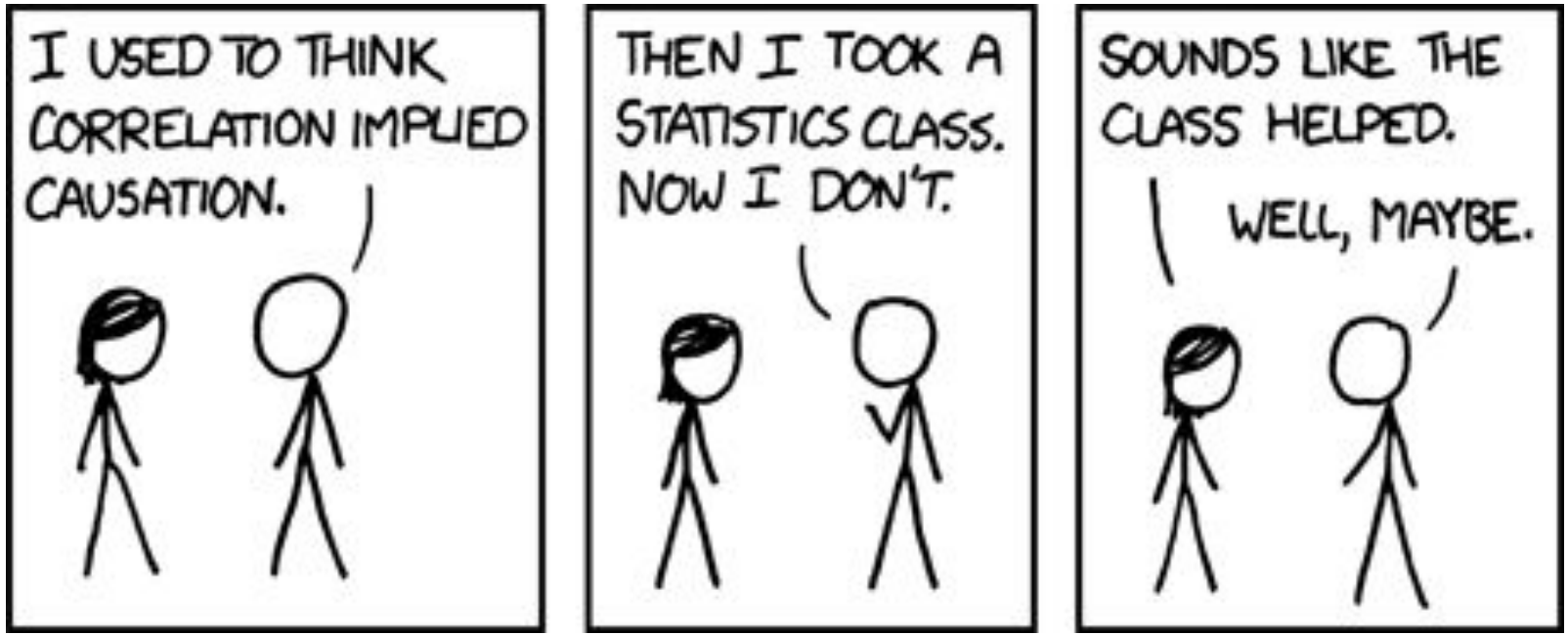
Correlation and Regression

- **Correlation:** Measures the strength of association between variables

- **Correlation Coefficient:** Expresses the strength of association between variables

- **Regression:** Predicts a value for one variable given the value of another variable

Pearson Correlations for Networks

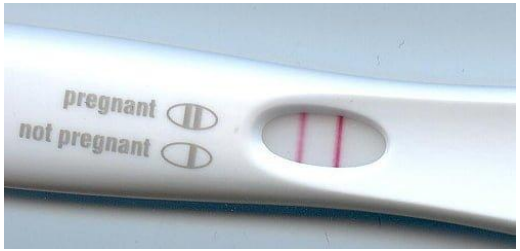


- Useful to know, despite not implying causation!

Img Src

Correlation does not mean causation

- Is a person's gender correlated with pregnancy?



Being female does not cause one to get pregnant!

Two variables (gender and pregnancy) are strongly correlated

Pearson Correlations for Networks

- Pearson correlations are often used to summarize pairwise structural equivalence
- Particularly useful when the data on ties are "valued"
  - It can tell us about the strength and direction of association, rather than simple presence or absence.

- Pearson correlation values range **from -1.00 to +1.00**
  - **-1.0 :** The two actors have exactly the opposite ties to each other actor
  - **+1.0:** The two actors always have exactly the same tie to other actors - perfect structural equivalence).

Img Src

Pearson Correlations for Networks

- Which of the following values indicate the stronger relationship (measured by Pearson's coefficient)?
  a. -0.20
  b. -0.68
  c. 0.5
  d. 0.12

Img Src

Pearson Correlation Coefficient Formulation

- Given paired data {(x_1, y_1), (x_2, y_2), … (x_n, y_n)} [In our case, these will be the entries in the adjacency matrix]

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

- N = Sample size (the number of entries in the adjacency matrices)
- x_i, y_i = Individual entries
- $\overline{x}$ = Mean of entries (analogously for $\overline{y}$)

Example for Pearson Correlations for a network

Figure 13.4. Pearson correlations of rows (sending) for Knoke information network

	1	2	3	4	5	6	7	8	9	10
1	1.000	0.447	-0.000	0.775	0.293	0.258	0.467	0.775	1.000	0.500
2	0.447	1.000	-0.447	0.447	0.655	0.293	0.333	0.745	0.333	0.378
3	-0.000	-0.447	1.000	0.258	-0.293	-0.149	0.600	-0.333	0.447	0.258
4	0.775	0.447	0.258	1.000	0.293	-0.258	0.745	0.775	0.775	0.775
5	0.293	0.655	-0.293	0.293	1.000	0.000	0.218	0.488	0.218	0.378
6	0.258	0.293	-0.149	-0.258	0.000	1.000	-0.447	-0.149	0.149	0.067
7	0.467	0.333	0.600	0.745	0.218	-0.447	1.000	0.600	0.745	0.258
8	0.775	0.745	-0.333	0.775	0.488	-0.149	0.600	1.000	0.600	0.149
9	1.000	0.333	0.447	0.775	0.218	0.149	0.745	0.600	1.000	0.600
10	0.500	0.378	0.258	0.775	0.378	0.067	0.258	0.149	0.600	1.000

- Strong correlation between ties node-1 and node-9 have
- Moderate tendency for node-6 to have ties node-7 does not have

Img Src

QAP - Example

Dependent

Row/ Col	1	2	3	4
1	Y1,1	Y1,2	Y1,3	Y1,4
2	Y2,1	Y2,2	Y2,3	Y2,4
3	Y3,1	Y3,2	Y3,3	Y3,4
4	Y4,1	Y4,2	Y4,3	Y4,4

Independent

Row/ Col	1	2	3	4
1	X1,1	X1,2	X1,3	X1,4
2	X2,1	X2,2	X2,3	X2,4
3	X3,1	X3,2	X3,3	X3,4
4	X4,1	X4,2	X4,3	X4,4

QAP - Example of a Permuted Matrix

Original Dependent

Row/ Col	1	2	3	4
1	Y1,1	Y1,2	Y1,3	Y1,4
2	Y2,1	Y2,2	Y2,3	Y2,4
3	Y3,1	Y3,2	Y3,3	Y3,4
4	Y4,1	Y4,2	Y4,3	Y4,4

1 ← 3
2 ← 2
3 ← 4
4 ← 1

Permuted Dependent

Row/ Col	1	2	3	4
1	Y3,3	Y3,2	Y3,4	Y3,1
2	Y2,3	Y2,2	Y2,4	Y2,1
3	Y4,3	Y4,2	Y4,4	Y4,1
4	Y1,3	Y1,2	Y1,4	Y1,1

Using QAP in practice

- QAP is designed as a bivariate test (only two variables).
- The networks need to be of the same size to be comparable.
  - Independent Network - X
  - Dependent Network - Y
- The larger the network, the higher number of runs expected for the same level of coverage
- In how many of the saved runs, was the observed ≥ new?
  - That’s the significance

You can run QAP on network data, attributes or metrics

- A network directly, such as an Agent x Agent network.
- A vector of node-level numeric attributes (e.g. age) repeated by row or column to form a network
- A vector of node-level measure values (e.g. Betweeenness Centrality) repeated by row or column to form a network.

QAP - Each Dyad is an Observation

Person	A	B	C	D	E
A	.	0	2	3	1
B	4	.	8	10	6
C	5	5	.	5	5
D	2	8	7	.	3
E	2	4	3	5	.

Pair	Row Number	Column Number	Absolute value of age difference	Friendship Rating
AA	1	1	.	.
AB	1	2	5	0
AC	1	3	25	2
AD	1	4	35	3
AE	1	5	15	1
BA	2	1	5	4

Src

QAP / MRQAP in ORA

Florentine Families

- Yet another famous benchmark dataset
- Models ultra wealthy Renaissance era Italian families' marriage and business ties.
- Extracted from historical documents (more info)
- You can run QAP to see if business and marriage ties are correlated

Src

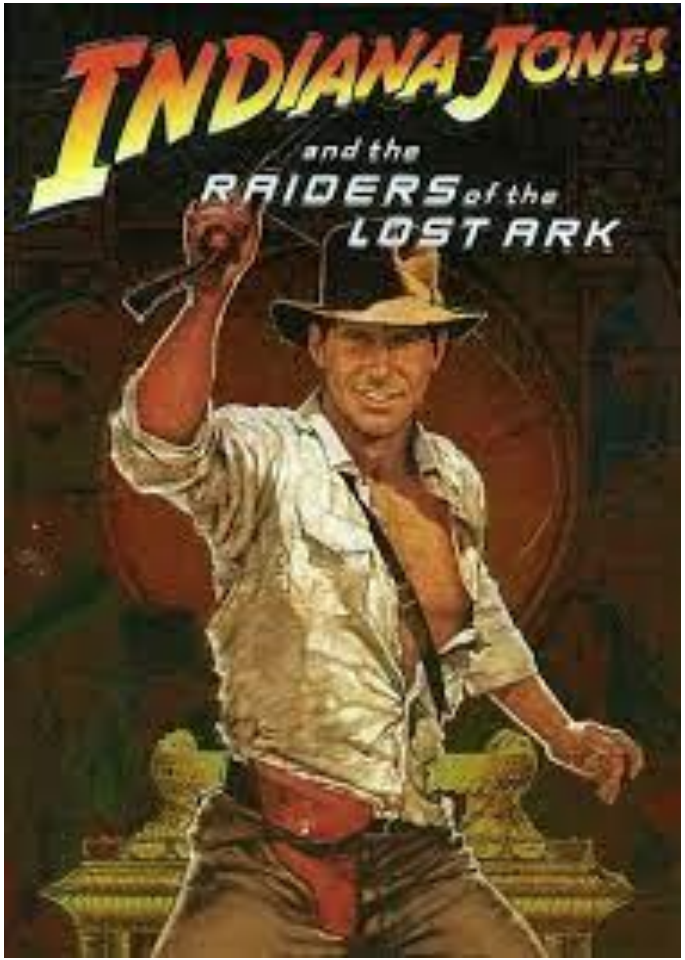
What if you have more than 2 matrices?

- Not a bivariate problem anymore, it is multivariate.

- Use MRQAP (**M**ulti **R**egression **QAP**) !
  - You have more than one independent matrix
  - You need to permute all of the independent matrices in the same way each iteration

Raiders of the Lost Ark

- Based on the movie Raiders of the Lost Ark (1981), directed by Steven Spielberg.
- Encodes the location of the characters for each time interval (where the location of characters is known).
- All meta-networks contain the same number of nodesets and nodes
    - 756 nodes in total are contained in the dynamic meta-network as a whole.



Available from: http://www.casos.cs.cmu.edu/tools/datasets/internal/index.php.

Dynamic Network Analysis in Gephi

Hospital Contact Dataset

- Dataset from “Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors“ (paper)
- What counts as contact ?
  - Wearing Proximity Sensors
  - Within 1.5 meters for 20 seconds
  - Patient, Nurse, Dr, Adm
  - 46 staff, 29 patients over 4 days 4 nights
  - 14,037 contacts were recorded

Infection Transmission Routes Network

- Dataset from SocioPatterns paper "What's in a crowd? Analysis of face-to-face behavioral networks" (link to data)
- **Nodes:** Visitors of Science Gallery
- **Edges:** Exists for contacts within face-to-face distance. Edge weight shows the number of contacts.
- One GML file for each of the 69 days
- Ids are reused across days
  - This is despite each visitor showing up only one day
  - This also means it is not a very accurate dataset although it simplifies some things

Infection Transmission Routes Network

- **High Level Steps**
  - Convert each gml file for the day to gexf format (Open in Gephi in gml format, export to gexf format)
  - Edit the gexf xml file to make it dynamic
    - **Before:** <graph defaultedgetype="undirected" **mode="static"**>
    - **After:** <graph defaultedgetype="undirected" **mode="slice" timerepresentation="timestamp" timestamp="1"**>