5-Weeks Course on Interactive Visual Network Exploration

Week1: Network Data Preparation

Jan 12th, 2022

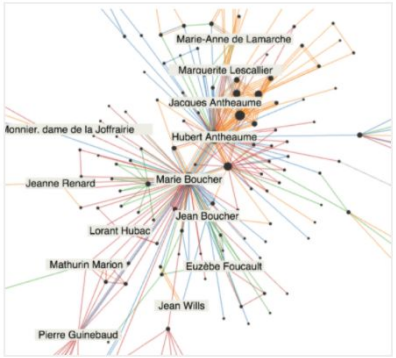
VIS HUB University of Edinburgh Visual+ Interactive Data

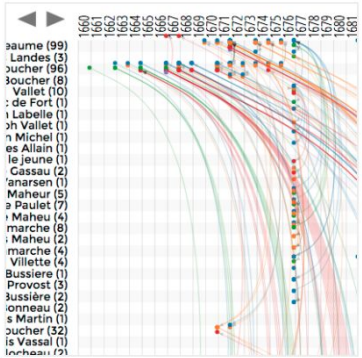
Inria

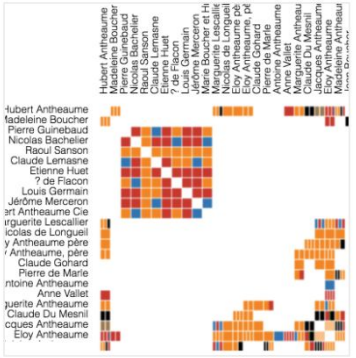
THE UNIVERSITY of EDINBURGH

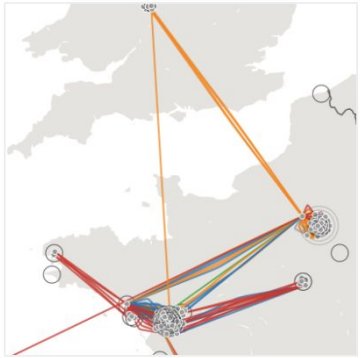
About the Course

- This free online course teaches the basics of exploring temporal, geographical, and multivariate network data (e.g., social networks) through interactive visualisations in The Vistorian (http://vistorian.net).
- This course will not assist you only in your visual network exploration but also in your network data preparation.



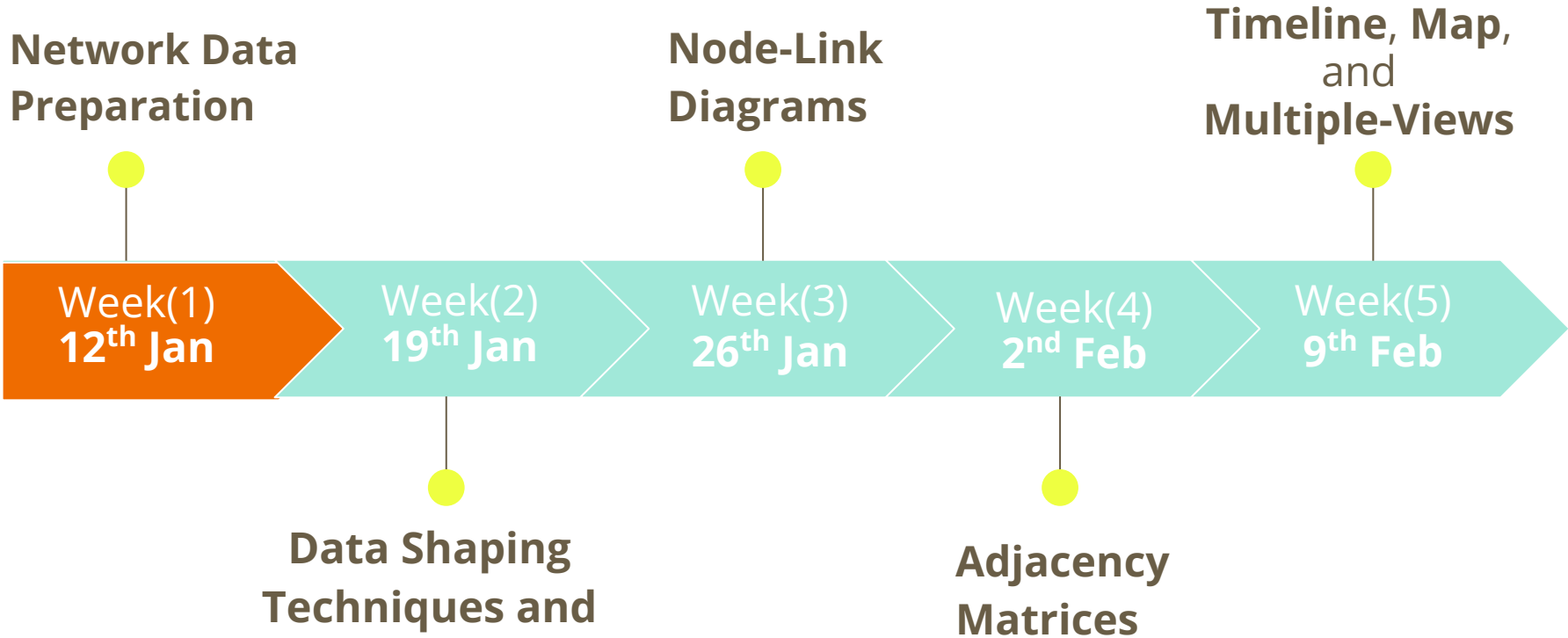






Course Goals

1. **Structure your network data and prepare it** for visualization with the Vistorian.
2. **Define goals of your exploration** and what you aim to learn about your network data using visualizations.
3. **Know a range of network visualizations**, through theory and hands-on use.
4. **Use different types of interactive visualizations** to explore your data.


Network Data Preparation
Node-Link Diagrams
Timeline, Map, and Multiple-Views
Week(1) 12th Jan
Week(2) 19th Jan
Week(3) 26th Jan
Week(4) 2nd Feb
Week(5) 9th Feb
Data Shaping Techniques and Challenges
Adjacency Matrices

Session 1:
Network Data Preparation

Jan 12th, 2022

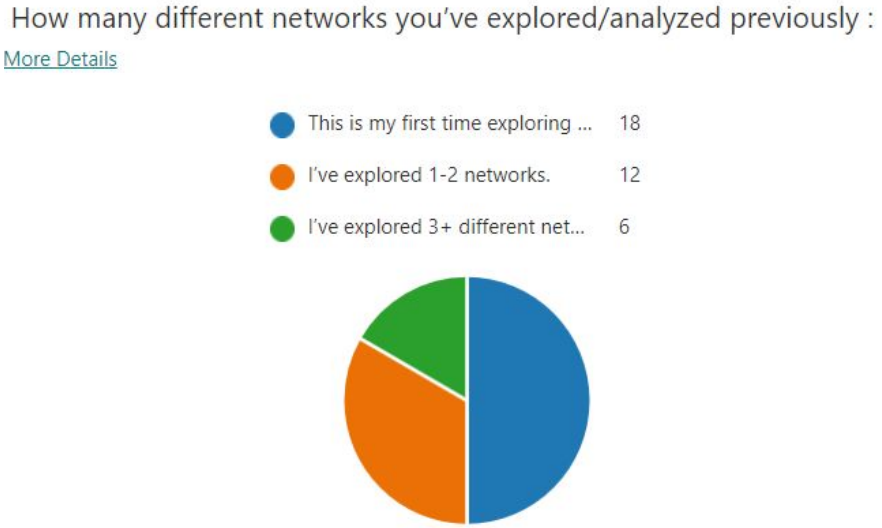
Mashael AlKadi

Session Outline

- Dataset selected for Hands-on Activities
- Ethics & Data Privacy
- What is a Network?
- What is Network Exploration?
- Role of Visualizations in Network Exploration
- Sketching your network
- Network Concept Map
- **Creating Node and/or Link Tables**
- **How can you format your network data?**

Insights

Network Exploration Experience	%	Have their own dataset
This is my first time exploring	**50%**	**Yes 45%** No 55%
I’ve explored 1-2 networks	**33.3 %**	**Yes 41.7%** No 58.3 %
I’ve explored 3+ networks	**16.7%**	**Yes 50%** No 50%



For experts

- Check the slides
- Follow along but feel free to start formatting your data
- https://vistorian.net/formattingdata.html

Activities Groups

History

Medicine & Health

Law & Politics

Business & Management

Natural Sciences

Diverse-Topics

Questions and Discussions During the Session

- Please poost your questions in our course's **Slack channel** https://join.slack.com/t/vistorian/shared_invite/zt-zo1w94tf-dkrbkqsRQqOTxavI3R~m_w

- Our website: https://vistorian.net/

Dataset selected for Hands-on Activities

Sender	Sender Location	Receiver	Receiver Location	Amount	Year	Type
Bob	Rome	Charles	Lisbon	10	1801	Loan
Bob	Paris	Charles	Lisbon	14	1803	Gift
Bob	Rome	Charles	Lisbon	3	1810	Purchase
Bob	Rome	Anton	London	2	1801	Purchase
Anton	London	Bob	London	5	1810	Loan
…	…	…	…	…	…	…

Your **Dataset** for Course Activities

- We expect participants to be working with their own network data during as the sessions will help you working with that data set.
- If you do not have your own dataset, You can use the Marie Boucher demo dataset: https://vistorian.net/gettingstarted.html .
- An ideal size for the course is up to 500 links in your network. Ideally, your network has fewer links to make best use of the activities and visualisations.

Your **Dataset** for Course Activities

- Your data can include data types such as:
  - **Categorical variables**: a variable that can take on one of a set of possible value (e.g. link type).
  - **Quantitative variables**: numerical values (e.g. link weight, location coordinates).
  - **Strings and text variables**: descriptive textual values (e.g. as node's label)

Your **Dataset** for Course Activities

Your dataset can *contain* any of the following aspects:

- **Geographic information** (such as place names). No coordinates are required.
- **Temporal information** such as the precise time a relationship is active (e.g., a day) or their start and end date.
- **Multiple links** between the same two nodes (such as individual letters sent between two people
- **Directed links** (such as a letter send from person A to person B) or undirected links
- **Weighted links** (such as strength of a social relationship)
- **Different types** of **links** (such as different types of social relationships)
- **Different types** of **nodes**.

Your **Dataset** for Course Activities

For my dataset file, In Which format should it be?

- Your dataset can be raw, formatted, i.e., not yet ready.
- Network files can be in any format such as:
    - **Tabular** data format (e.g., a spreadsheet or CSV file)
    - **Network** format (e.g., GEXF, GraphML, GXL) (S*oon to be supported*)

Ethics & Data Privacy

- Please ensure that you don't share any private data that cannot be disclosed with others.

- This course is part of a research project which was approved by the Ethics Board of the School of Informatics Approval #2019/67905

- All approved ethics-related documents can be found at: http://vistorian.net/courses including:
  - Participant Information Sheet
  - Participant Consent Form

The Vistorian Team

Researchers

- New visualizations
- Training
- Support workflow

Users

- Issues with vis
- Real-world problems
- Feedback

Week (1): Network Data Preparation

Sender	Sender Location	Receiver	Receiver Location	Amount	Year	Type
Bob	Rome	Charles	Lisbon	10	1801	Loan
Bob	Paris	Charles	Lisbon	14	1803	Gift
Bob	Rome	Charles	Lisbon	3	1810	Purchase
Bob	Rome	Anton	London	2	1801	Purchase
Anton	London	Bob	London	5	1810	Loan
…	…	…	…	…	…	…

What is a network?

- A network is a set of interconnected data elements that are known as a set of nodes and edges.
- Most characteristics of networks originate from graph theory.

“Graph” vs. “Network”

Graph

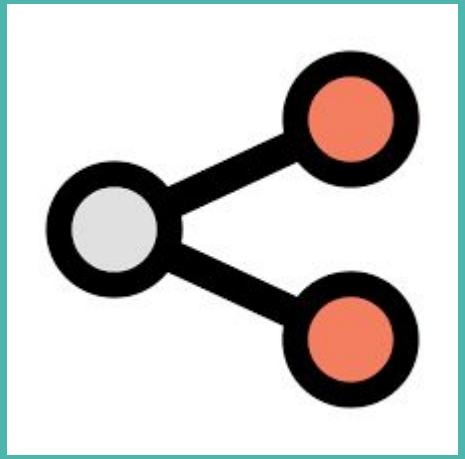
- Usually used to refer to the mathematical and topological concept: nodes & links.
- Usually used to refer to the graph-*layout*

Network

- Usually used to refer to graphs where nodes and links have additional data: geography, time, types, directionality, et...

It’s fine to use both terms interchangeably.
But **‘graph’** is also often used to talk about **‘charts’** or **‘visualizations’** in general. **Network** is more clear and specific.

Networks Fundamentals



What is a Node?

- A node (also known as actor, vertex, point, entity ...)
- Nodes are usually perceived as the **ontological entities** who engage in relations / are related.
- Nodes can be individuals, groups, organizations, locations, events, concepts, ideas, words, a data entry, ...
- Nodes can have properties, which are key-value pairs (e.g. node permanent location, node type, node size, ...).

Point
Actor
Vertex
Nodes

Link
Arc
Relation(ship)
Connection

What is a Link?

- A link (also known as a edge, tie, relation, ...) have exactly two endpoints (nodes)
- Edges are expressions of **ontological relations**: how are two nodes (entities) related?
- E.g., friendship, collaborations, trade exchanges, communications, similarity, distance, flow, requirement-for, etc...
- If the two endpoints are the same node then the link is called **Recursive/ Loop** relation.
- Each edge can have its own ***properties*** (e.g. time of creation, type , ..)

Terminology

Different terminologies may be used based on the domain.

During this course we will stick to specific terms: such as node and link.

Please don’t hesitate to ask for clarification at any point during the course.

Directed Link

Multiple Links

Link Types

Node type

Weighted link

Combinations

Link Direction

Directed:

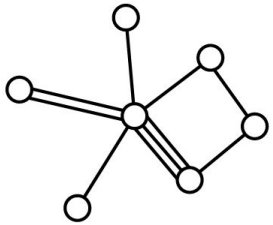
- The relationship has a direction.
- A directed edge (arc) is an ordered pair of nodes in which the first node the sender (the tail, source) and the second the receiver of the link or tie (the head, target).
- Directed Graph/Network or digraph
- Example: Emails and letters

Undirected:

- The two items are linked without the concept of direction; both individuals are equally involved in the relation.
- Undirected Graph/Network
- Example: A social relation that is undirected (e.g., cooperation on friendship)

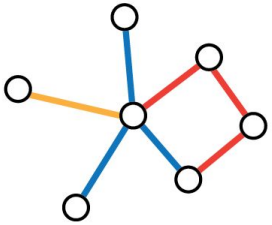
More about Networks

- In a graph, multiple links are allowed, but when we say that a **graph is simple**, we indicate that it has no multiple lines.
  - A simple undirected graph contains neither multiple edges nor loops.
- **Link weight**: indicate the strength of a relation, which is a quantity.
- **Link Type**: network can contain multiple types of links/relations. This creates a multiple relations network, which is also called a *multiplex network*.



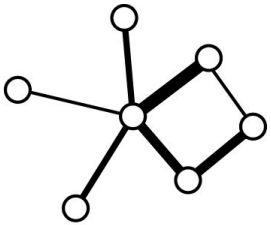
Multiple links

+



Link types

+

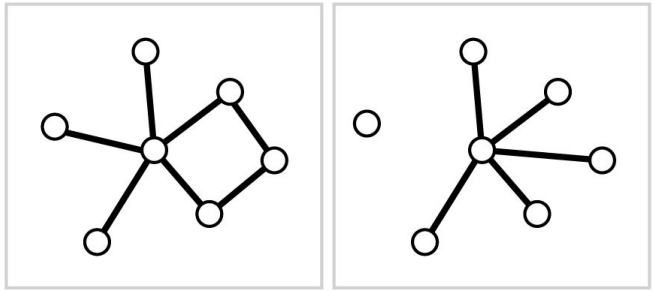


Link weight

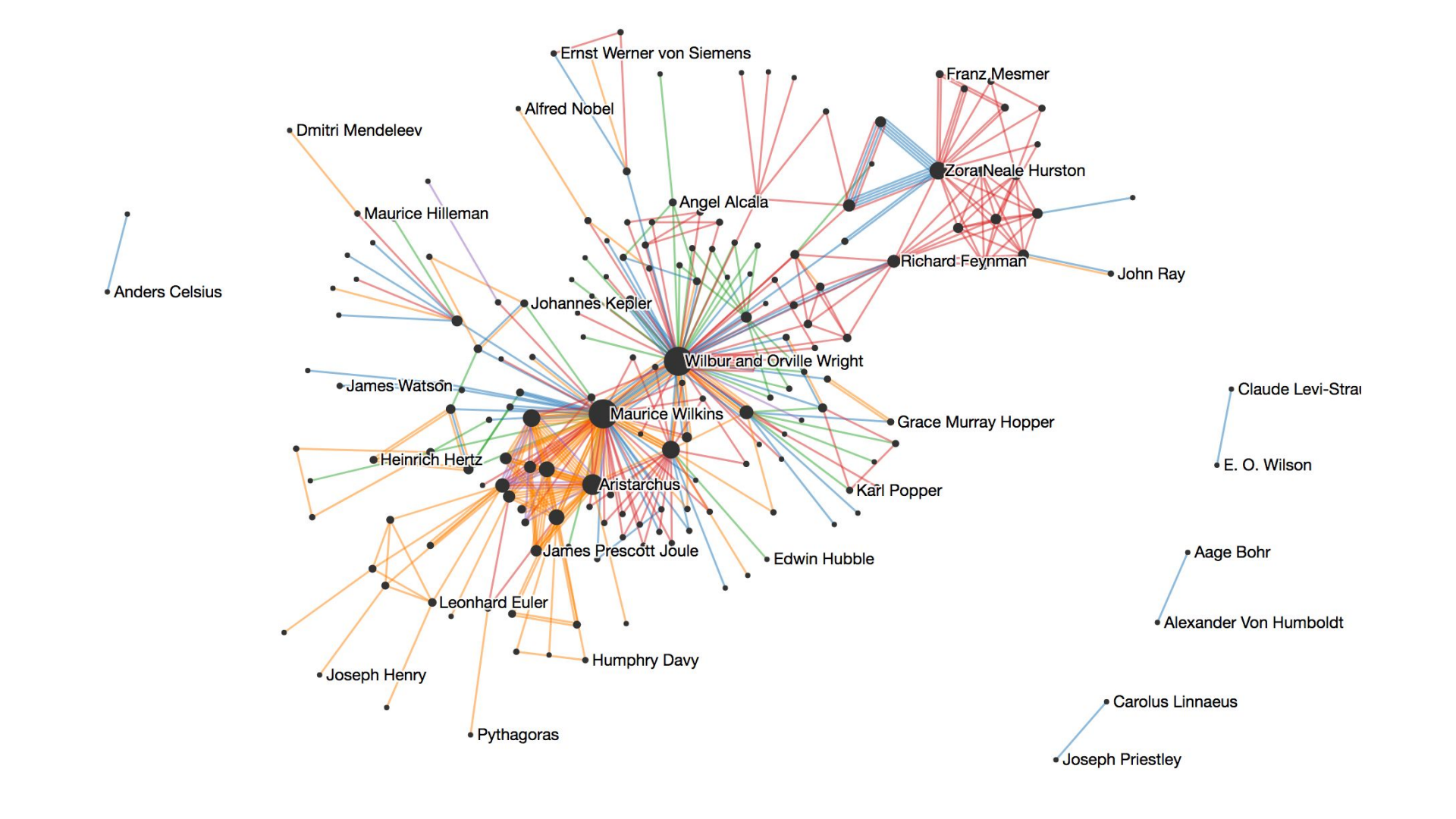
+

Geography

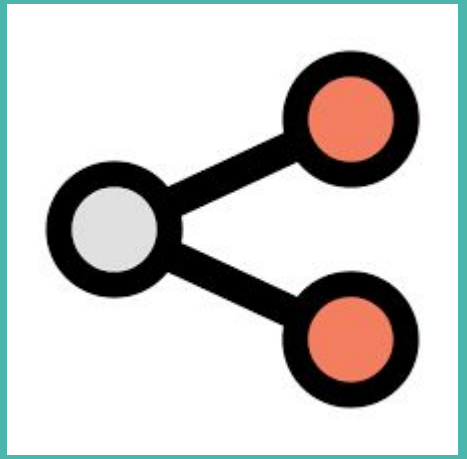
+



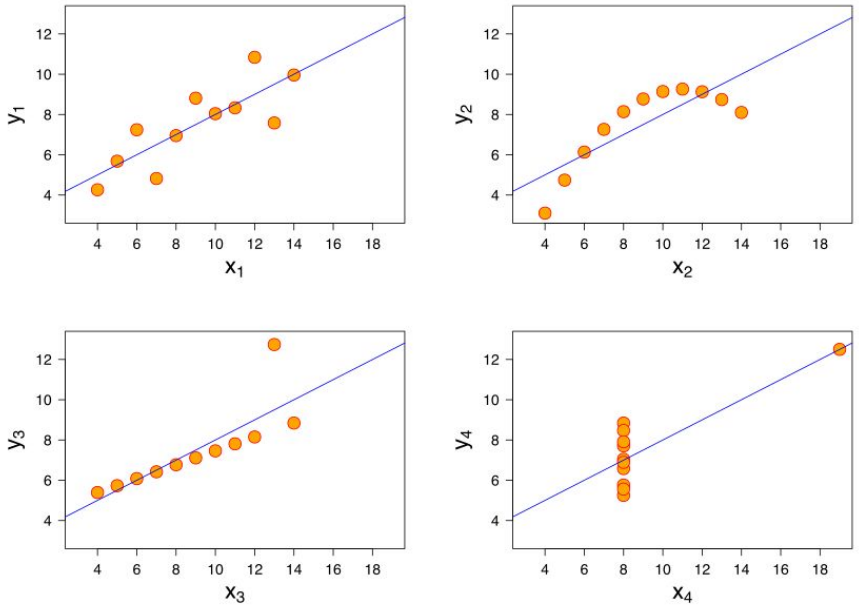
Changes over time


Ernst Werner von Siemens
Franz Mesmer
Alfred Nobel
Dmitri Mendeleev
Zora Neale Hurston
Angel Alcala
Maurice Hilleman
Richard Feynman
John Ray
Anders Celsius
Johannes Kepler
Wilbur and Orville Wright
James Watson
Maurice Wilkins
Grace Murray Hopper
Claude Levi-Strau
E. O. Wilson
Heinrich Hertz
Aristarchus
Karl Popper
James Prescott Joule
Edwin Hubble
Aage Bohr
Leonhard Euler
Alexander Von Humboldt
Humphry Davy
Joseph Henry
Carolus Linnaeus
Pythagoras
Joseph Priestley

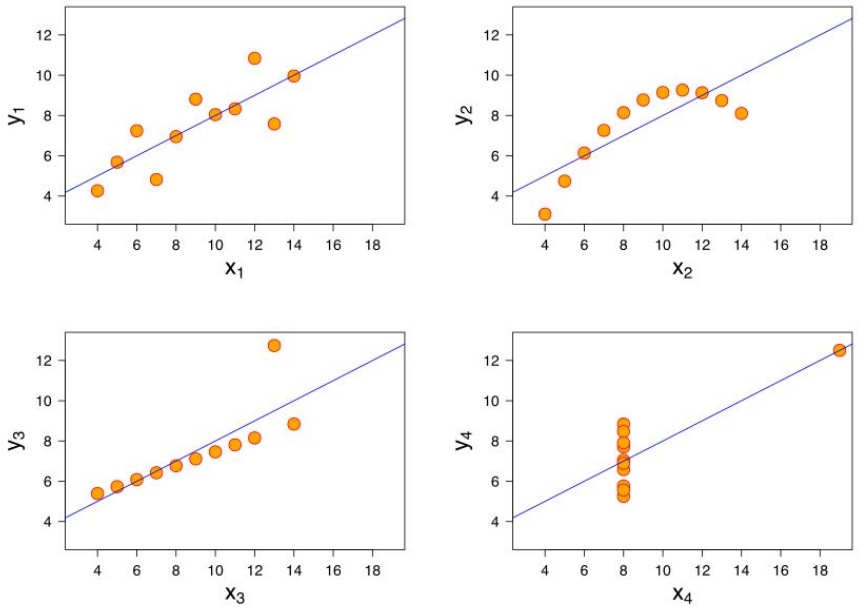
Network Exploration



Anscombe's Quartet



Anscombe's Quartet



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

What is Network Exploration?

- Exploratory network analysis consists of four parts :
  a) the definition of a network,
  b) network manipulation,
  c) determination of structural features,
  d) and visual inspection.

- Exploratory network analysis means that *we have no specific hypotheses* about the structure of a network beforehand that we can test.

What is Network Exploration?

John Tukey's concept of **"Exploratory Data Analysis : the primary purpose of visualizing and exploring is to raise questions and gather insights about a large quantity of data."**

Ben Shneiderman's mantra: "**Overview first, zoom and filter, then Details-on-demand"**

What is Network Exploration?

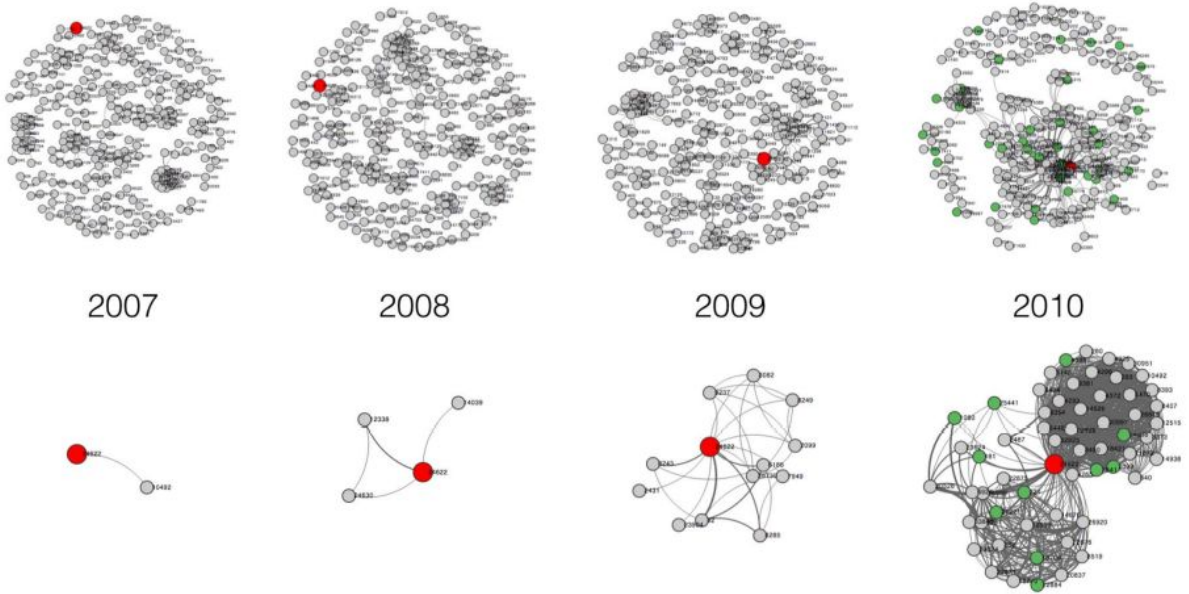
“Information visualization is sometimes described as a way to answer questions you didn’t know you had.”

“Discovery is seldom an instantaneous event, but requires studying and manipulating the data repetitively from multiple perspectives and possibly using multiple tools.”

Catherine Plaisant “*The Challenge of Information Visualization Evaluation* “

Networks and their **Role** in **Data Exploration** and **Analysis**

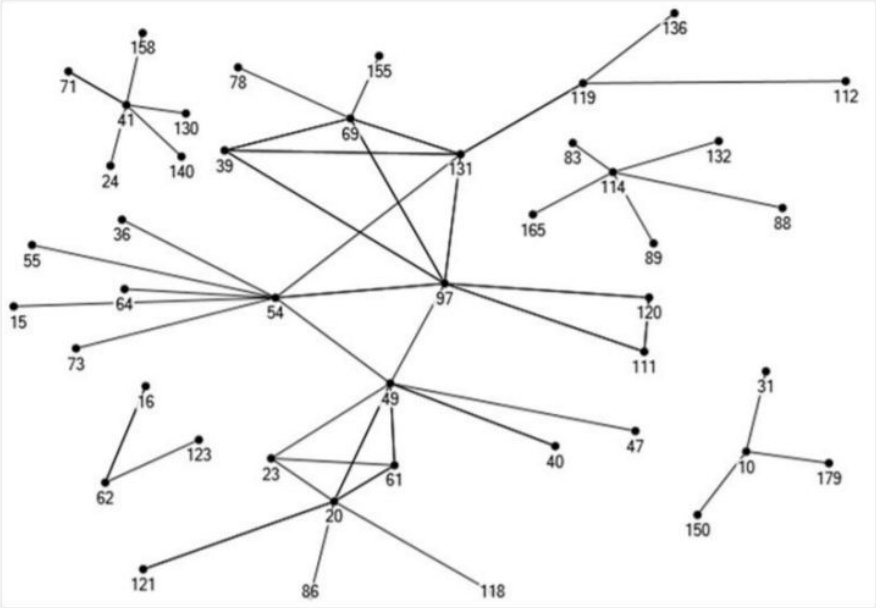
Role of Visualizations in Network Exploration



CollaborationViz: Interactive Visual Exploration of Biomedical Research Collaboration Networks by Bian et al.

Role of Visualizations in Network Exploration

Role	Betweeness (Unique links to others in the network)	Degree (Connected to many individuals)	Characteristics
Gatekeepers	↑ Higher	↓ Lower	• May play an important role in activity, but not much information is held on them • Removal may fragment networks
Highly visible figures	↓ Lower	↑ Higher	• May have information about many others in the network • May be involved in lots of activity in the network, but do not play a unique role
Central figures	↑ Higher	↑ Higher	• Very visible and central role • Key figures that may be focused on to fragment networks and to gather information



Social Network Analysis of local gang issues by the Home Office

Role of Visualizations in Network Exploration

- A network is only a tool, a lens, a perspective that helps analyse your data.
- A network represents entities and relationships about your data.
- For your network analysis, you first need to define those network(s) that can best help you answer or discover your research questions. We will exercise this later in the tutorial.

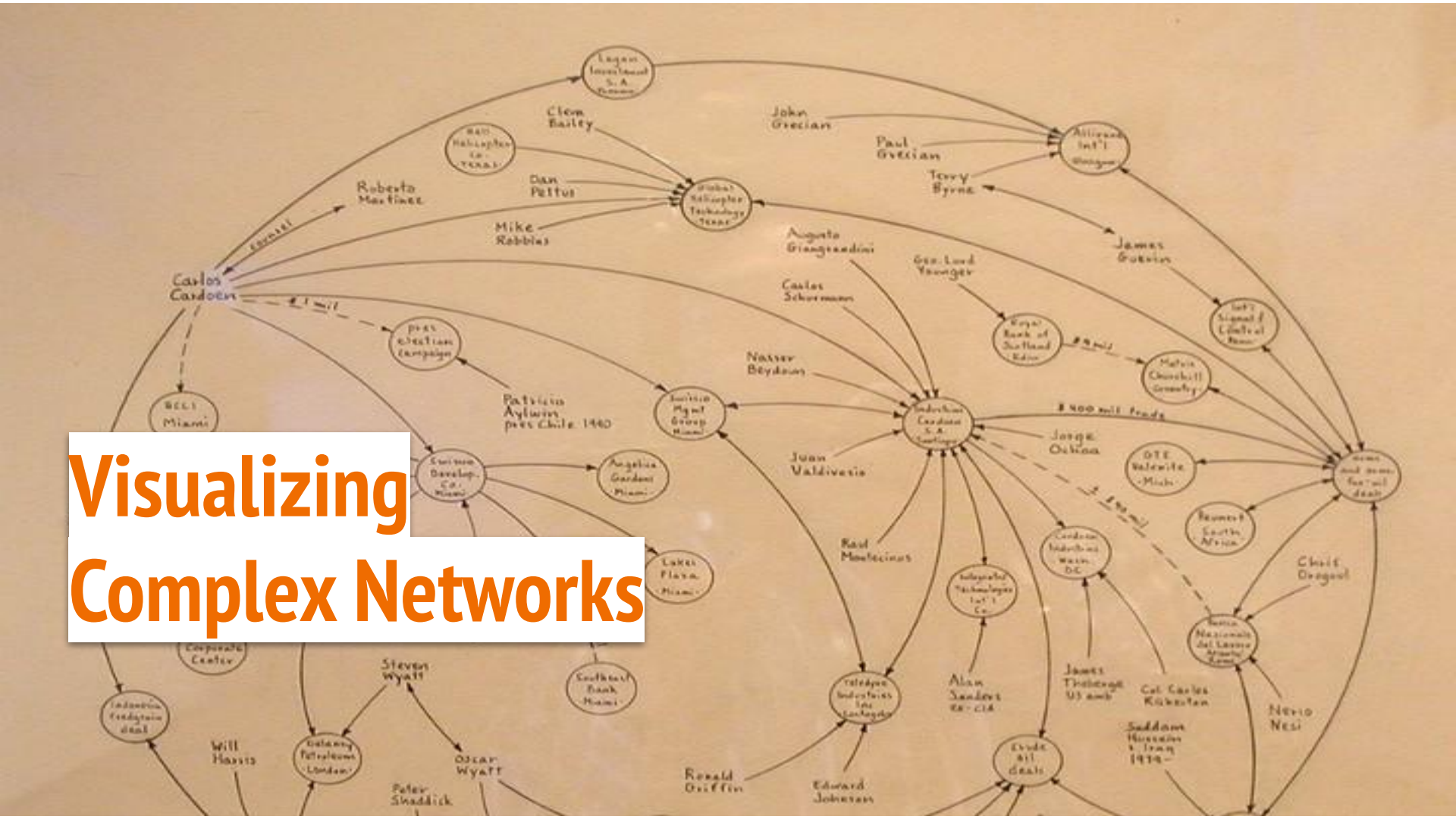
Role of Visualizations in Network Exploration

- **Depending on what you define as nodes and links, you can create potentially a high number of networks from your one data set.**
- For examples of networks created from the setting of academic publishing, we can create 3 different networks as:
  - (a) coauthorship network (person--person),
  - (b) papers with similar authors (paper--paper),
  - (c) person--paper

Exercises

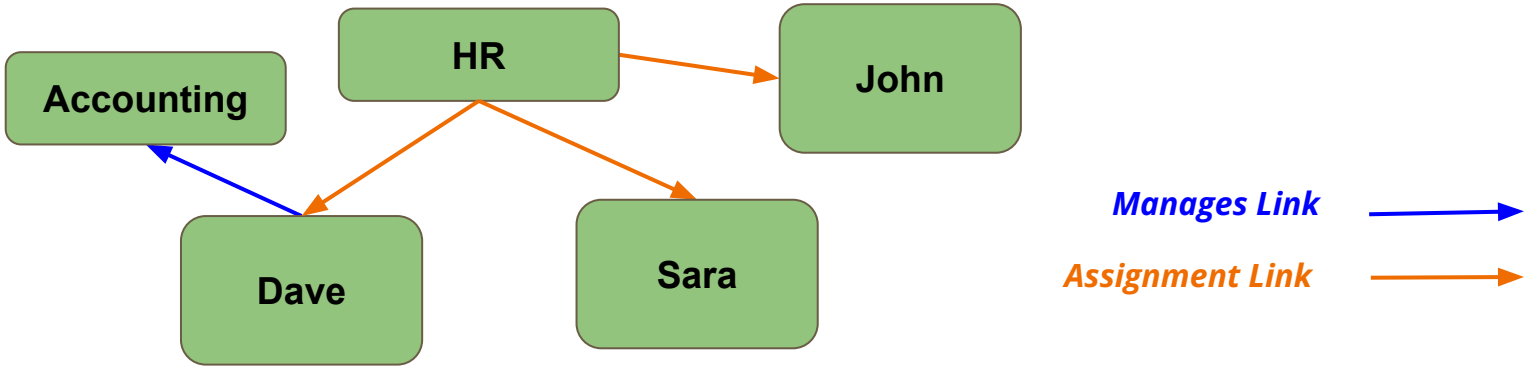
Ice-breaking Activity

- Please prepare a short **pitch** about (3min):
  - What is your **Data** about?
  - Do you have **research questions that** you want to answer from your network and data?


Visualizing
Complex Networks
Carlos Cardoen
Roberto Martinez
Clem Bailey
Dan Pettus
Mike Rabbitt
John Grecian
Paul Grecian
Terry Byrne
James Guerin
Augusto Giangreandini
Gen. Lord Younger
Carlos Schermann
Natser Beydoun
Patricio Aylwin
Juan Valdivieso
Jorge Ochoa
Raul Moncecinos
Alan Sanders
James Theberge
Col. Carlos Kirkerian
Nerio Nesi
Steven Wyatt
Will Harris
Oscar Wyatt
Peter Shaddick
Ronald Griffin
Edward Johnsen
Saddam Hussein
Iraq 1979-

Exercise (1) : Sample Sketch of your Network (3min)

- Draw an imaginary sketch of how your network would look by choosing 2-4 nodes only.
- Draw the possible links/edges between them.
- List a set of questions you think this network can help you answer.



Exercise (2) : Network Concepts (3-5min)

Nodes:

1. List possible nodes in your data. Nodes can be elements or concepts such as person, letter, research paper, .. etc.
2. How many different types of nodes can you identity in your network?
3. Which of your list elements can be selected as node(s) of interests?

Exercise (2) : Network Concepts (3-5min)

Links:

5. Now, explain how these nodes can be related to each other.
    a. In what type(s) of relation does a node (element) connect to another element?
    b. How many relations can a single node connect to another node?
6. How many different types of links can you identity in your network?
7. Which of those listed you are interested in selecting as edge/links?

Exercise (3) : Create Possible Networks (3 x 3-5min)

1. Pick a set of nodes and and links from previous exercise
2. Draw a small network (around 15 nodes) of **how you imagine** your network topology to look (you can use color, point size, labels, etc. to show additional information).
3. List down questions and possible information you could be able to observe.
4. *Repeat 3 times with different nodes and links.*

Exercise (4) : Concepts Map / Data model (5min)

Use a new A4 sheet so you have space for annotations later!

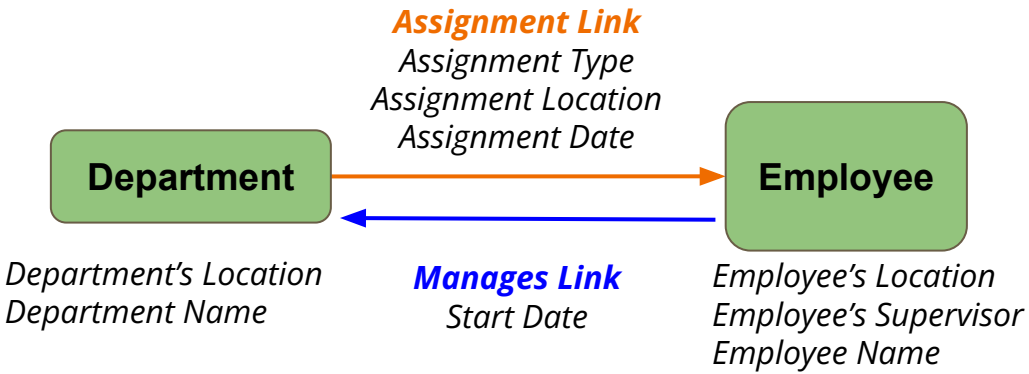
Exercise (4) : Concepts Map / Data model (5min)

5. For each distinct node type, draw a node shape/entity (ex. rectangle). Write inside the node's shape, the name of the node.
6. Determine the unique ID of your node. In other words, a unique value to distinguish between any two nodes.
7. For each distinct link type, draw a line between the involved nodes' shapes. Write on it the name of the link type.
    a. If it is a directed network : use an arrow to link nodes
    b. If it is a undirected network : use an line to link nodes
    c. If it is a relation between two nodes of the same type: draw a line from the node to itself (recursive).
8. **Congratulations**! Now you have your basic network layout.

Exercise (4) : Network Concepts Map

9. Adding additional (optional info) :
    a. List any properties that describe the nodes themselves: Add those properties to nodes node element (eg. location, type, .. etc)
    b. List any properties that describe the link between any two nodes: Add those properties to link line (eg. location, type, .. etc)
    c. How can I differentiate between adding properties to nodes or links?
        i. If the property describes a node in a persistent nature, then it is a node property.
        ii. If the property describes a node in a varying nature based on the link type, then it is a link property. (for example, the node location changes based on the relation type, or time, or location)

Exercise (4) : Network Concepts Map



Exercise (4) : Network Concepts Map

Refining your Map:

11. Can you think of any other nodes from different types that you can add and might help answer your questions; if yes go to step 1; else proceed to the next part.
12. In your current node and link lists, do you have any data that you can exclude from your network? That will not contribute to answering any of your questions.

Formatting Network Data into Tables

Types of Network Tables

1. Using **Link table** only
2. Using **Node table** only
3. Complementing **link table and node table** by each other.

Node Table

- Each **row** describes a **single NODE** and it should contains :a **node label** and at least **one relation**
- More descriptive data about the node can be added such as: **node location** and **shape** (differentiates between various types of nodes)

CHILD	MOTHER	FATHER	GOD-FATHER	GOD-MOTHER	PLACE-OF-BIRTH
Bob	Celine	Charles	Dave	Eve	Paris
Ana	Fannie	Gerd	Mike	Dianne	London
Celine	Maria	João	Pedro	Ana	Lisbon

Link Table

- Each row describes a single LINK between Source → Target nodes.
- More descriptive data about the link can be added such as: location, time, type, and weight

ID	Sender	Receiver	Money	Year
0	Anton	Bob	100	1801
1	Anton	Bob	30	1803
2	Anton	Charles	10	1801
3	Anton	Charles	20	1802
3	Anton	Charles	30	1803
4	Anton	Charles	100	1804

Complementing Link Table & Node Table by each other

Eventually, some networks may require both node and link tables:

- link tables to specify links and their attributes, and
- node tables to specify nodes and their attributes.

A node and a link tables are related *through node names*. I.e the node names in the node table must match the names of source and target nodes in the link table.

Exercise 5: Formatting Data

Exercise (5) : Creating your Network Tables

How can I decide the table type that I need?

1. **Link table**: if you have all of your data properties (attributes) placed on links only.
2. **Node table**: if you have your properties placed on nodes only.
3. **Both link and node tables**: if you have properties placed

Exercise (5) : Creating your Network Tables

Pick one of your networks from exercise 3

1. On paper, create a link table for that network (with fake entries if you do not know the real data)
2. If you have node attributes (type, etc..) create a node table for your network.
3. If there is still time, repeat for your other networks from exercise 3.

Wrap up

General Questions?

Next Week

- ***Need your tables***
- Next week's session:
- Ensuring Consistency of your Data
- Tools to assist in Checking Data Consistency
- Common Challenges in Network Visualizations such as:
  - Dealing with unstructured data.
  - Visualizing and exploring large data
- Starting your Exploration Plan
- Importing your Data to the Vistrian

Homework

- Create your tables (at least for a part of your data)
- Pass by our drop-in session for any questions : Monday 3-4 UK time or by email at m.alkadi@sms.ed.ac.uk
- Mini Feedback Form: https://forms.office.com/r/SS4vWNC028