Monday, May 12, Practical Session I: Data

Ryan Horne
Advanced Research Computing, UCLA
ryan.matthew.horne@gmail.com

Overview

- Data practices
  - Data preparation
- Discussion of network terms and concepts
  - Software demonstration (Gephi, Cytoscape, and Python oh my!)
  - Python is live!

Starting With Data

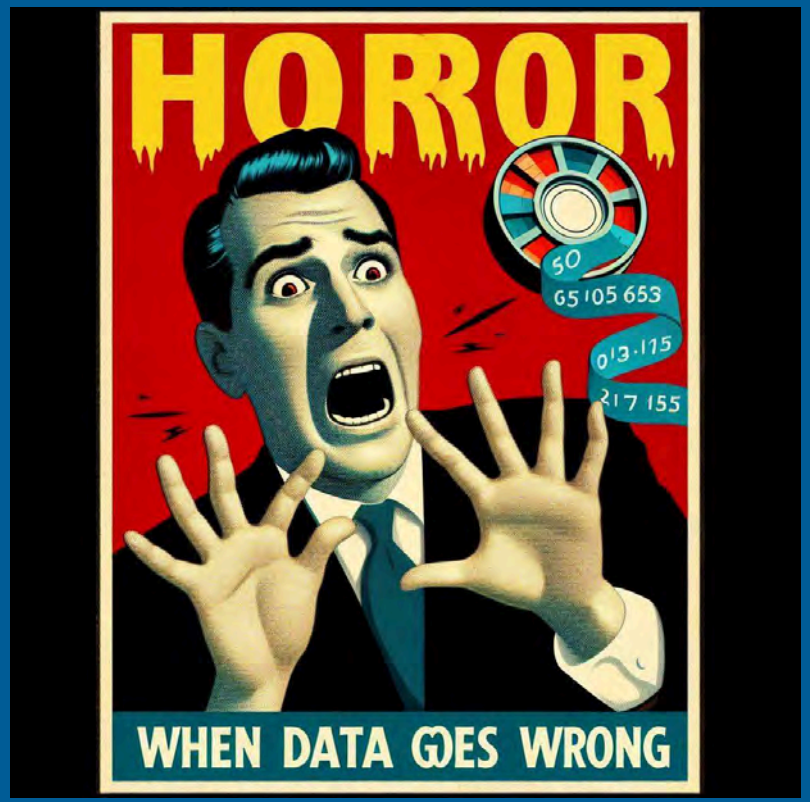
- What is data?
- How do we turn data into *information*?
- What rules do we follow?
- Is/are data a cultural construct?

Basic point: We have data, and even two people working on the exact same source with the exact same questions will *probably* have different data needs.

(Open) Data Is Everywhere

- GitHub, Zenodo, Dataverse, Dryad, etc etc
  - Often a requirement for funding
    - At least it *was*
- Linked Open Data (LOD)
  - Different data sets sharing common links
- Option for rich metadata
  - Data that describes data
- Are we in a data glut?
  - Research data
  - Personal data
  - Government data

When Data Goes Wrong



- Generated by an AI
- Notice anything?

When Data Goes Wrong

Humanities Data

- Most information is either *unstructured* or arranged in an individually idiosyncratic way
- Unable to be directly fed into standard tools and methods
- How do we avoid data catastrophe?
  - Technologies such as Linked Open Data and adherence to FAIR practices helps

FAIR and CARE Bears

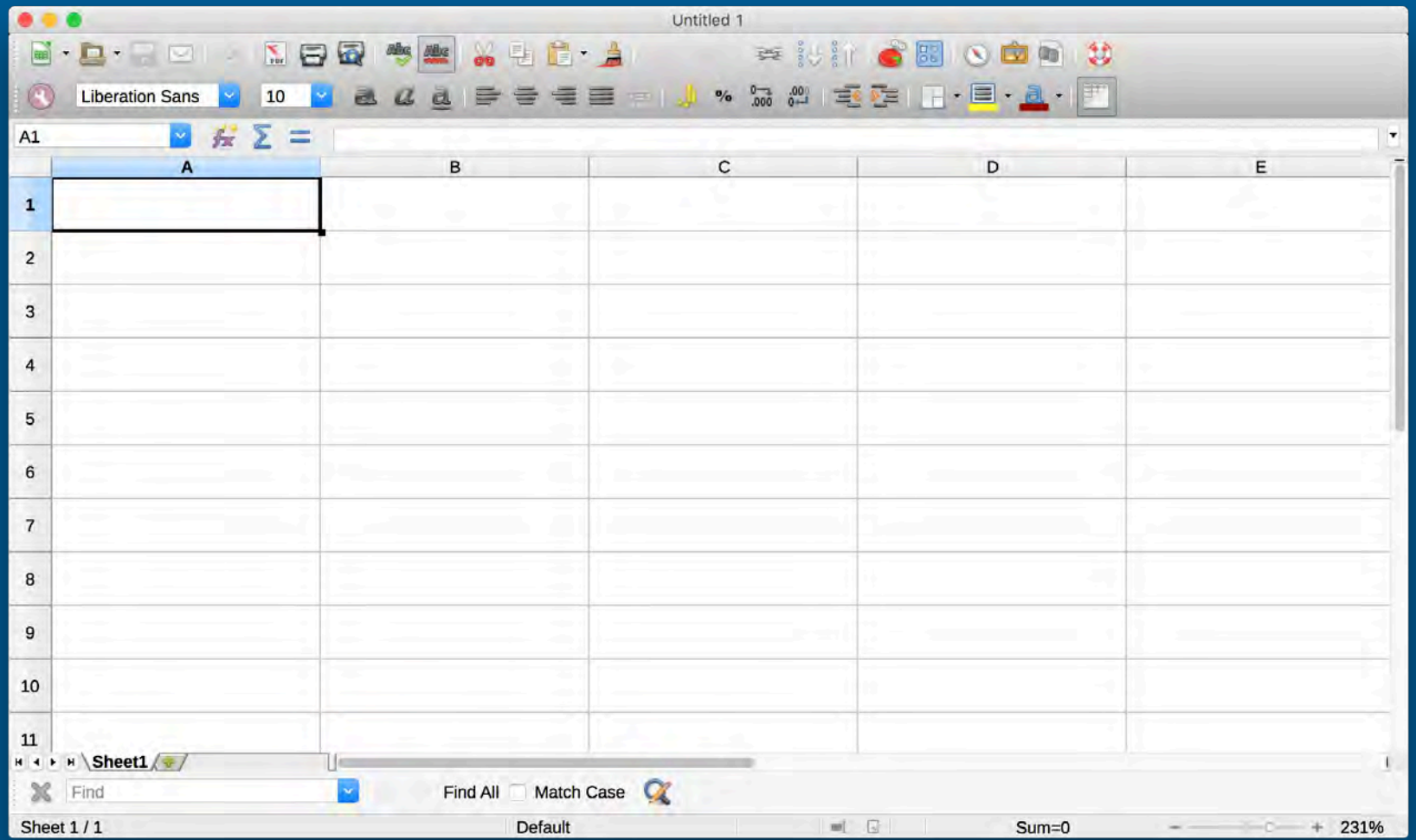
- FAIR data is:
  - Findable
  - Accessable
  - Ineeropreable
  - Reusable
- CARE is:
  - Collective benefit
  - Authority to control
  - Responsibility
  - Ethics

Not just communicate the data, but the process and the modeling structure of the data itself, with the assumptions, choices, ideas, and methods

Some Things to Consider

- What is your data?
- What are you trying to show?
- How are you going to organize your data?
  - Are there standards in your field?
- In short: Think a lot about database design!

Data Tools: The Basics



Basics

- Most basic level: Spreadsheet with two columns
- Entries need to be in the same format AND have the same spellings, etc
- AT MOST two different *types* of entities
- More information is good!

Some Best Practices

- Create a unique identifier
- Each column contains the same category of data in every row in the column
- Each row in the spreadsheet contains all of the fields of data for one entity
- The first row of the spreadsheet must contain a unique name at the top of each column
- No blank rows
- Think about attributes!

More Best Practices

- Do not be afraid to make multiple spreadsheets!
  - These can translate into multiple tables in a database
- Think about the ***minimum*** work needed to maintain and update your data
- Document!

More Best Practices

- Save. Save all the time. Save whenever you do something.
- Document, document, document!
  - Think of metadata and readme files as a love letter to yourself
  - Organization is key; although it might change
- Backup!
  - If you have not, get a free dropbox account NOW. Come back when it is done :)
- For further exploration: Version control
  - GitHub, etc.
  - More complex than what we have been talking about, but worth it if you are working with a group or are changing things all the time

Avoiding issues: Preparing your data

- DOCUMENT, DOCUMENT, DOCUMENT!!!!!
  - Every file, column, and data type should be in a readme.txt
  - Not just for others; also for you. Can you remember every shortcut / abbreviation you made? What about 6 months from now? A year?
- DOCUMENT all cleaning decisions
  - Are you rejecting a measurement? Why?
  - Is some source / text not relevant? Why not?

Testing your data

- Check columns, entries, etc for correct types
- Spot check random entries
- Look for outliers in your data set
- *Work* with your data. The more you hit the data set, the more issues you may find
- Does the data make sense?
- Be aware of tools that may introduce errors or smooth over your errors (null fields, encoding, etc)
- DOCUMENT, DOCUMENT, DOCUMENT!

Data Quality Control and Assurance

- Think about New York City.
- Will a computer recognize New York?
  - NYC?
  - New York, New York?
  - The Big Apple?
  - The City?
  - A zip code?
  - Spatial coordinates?
  - What about spelling or pronunciation issues (new yark)?
- Data cleaning may be needed

Data Cleaning

Data Cleaning can be a long process

- Cleaning almost always requires scholarly interpretation
  - Spelling / name issues
  - Matches to different data sets
  - Disambiguation
- What are your sources of error? How do you address these in publications?
  - Do you offer “raw” data ***and*** “clean data”?
  - How do you avoid issues with the public?
  - Is cleaning data misrepresentation?

Data Analysis Questions

Your data is in order! Congrats! Now…

- What are your units?
  - What do you have on each axis? Where do they start?
- Are you being representative?
  - Are you minimizing something?
  - Are you putting undue emphasis on something?
- Your networks ***are an argument***, just like any other publication or analysis.
  - There is no such thing as a “neutral” or “objective” data visualization or model!

Data Modeling

- Underpinning any network analysis or digital information system is the creation and classification of data itself
- And systems to present and publish it
- The choice of tools and models are intertwined
- Essential step in the analysis and use of data
- All data is a representation and simplification

Data Standards

- Data standards are an effort to ensure that data remains interoperable across different systems and tools.
- Common data standards in the humanities
  - RDF
  - XML
  - TEI (Text Encoding Initiative)
- Graph data models
  - RDF triples
  - Linked (Open) Data / Semantic web

Vocabularies

- One of the first steps in any data model is identifying the entities (or classes) in the model, and standardizing their spelling and definition
- Creating *metadata*
  - Data about the data
- Results in the creation of a *vocabulary*
  - Terms and definitions

Vocabulary Example

Term	Definition
Tweet	A message posted on Twitter, limited to 280 characters.
Hashtag	A keyword or phrase preceded by # that categorizes tweets.
Retweet	A repost of another user's tweet.
Protest	A public demonstration of objection, often against policies or actions.
Activist	An individual advocating for political or social change.
Pipeline	A structure used for transporting oil or gas, often a focus of environmental protests.
Water Protector	A term used by Indigenous activists opposing projects that threaten water sources.

Example Vocabularies

Vocabulary	Scope	Link
TEI (Text Encoding Initiative)	Markup standard for representing texts in digital form	TEI
EpiDoc	Subset of TEI for epigraphic, papyrological, and manuscript materials	EpiDoc
LCSH (Library of Congress Subject Headings)	Controlled vocabulary for subjects in library catalogs	LCSH
AAT (Art & Architecture Thesaurus)	Vocabulary of terms for describing art, architecture, and material culture	AAT
TGN (Getty Thesaurus of Geographic Names)	Hierarchical vocabulary for place names	TGN
ULAN (Union List of Artist Names)	Standardized names and biographical data for artists and creators	ULAN
Europena Eagle Vocabularies	Controlled vocabulary used for material culture (emphasis on epigraphy)	EEV

Ontologies

- Vocabulary masks the semantic structure of content
- Expression of the relationships / connectivity of a domain is an *ontology*
  - Hold the information about the possible values, limits, constraints, and relationships between those conceptions

Ontology Rules (Kind of)

Ontologies are built around the formal structures of

Classes

Properties

Rules

- Classes conventionally written with an initial uppercase letter
- All lowercase letters for properties

Classes

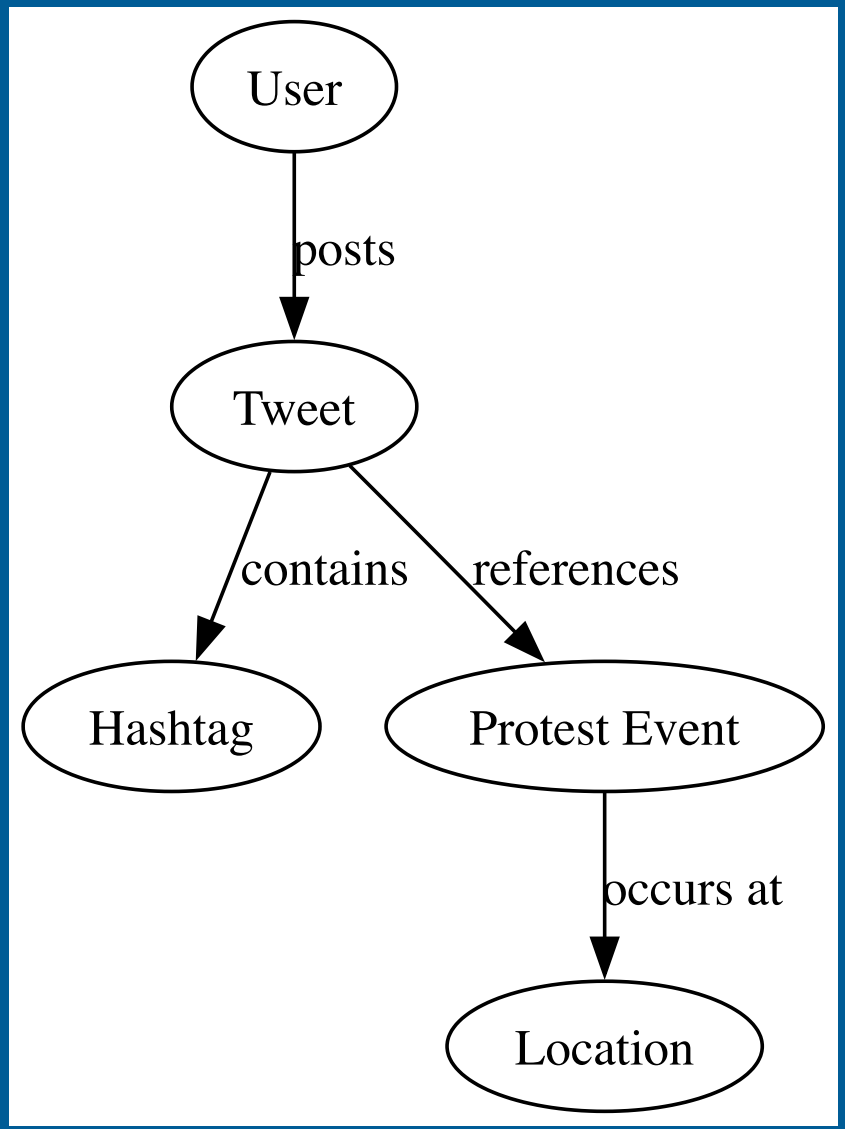
Properties

- Characteristics of *classes*
- A Town could have
  - A *population* property
  - A *location* property
  - etc
- Think of the columns on a spreadsheet

Rules

- Define how classes can interact
- Also the potential values they can hold
- A Town could be connected to
  - A Statistical Area or
  - A Ghost Town class
  - And not both
- A dates property could be configured to accept only numbers, etc

Example Ontology



Example Ontologies

Ontology	Scope	Link
CIDOC CRM	Conceptual model for cultural heritage data	CIDOC CRM
FOAF	Describes people and their social relationships	FOAF
Bibframe	Bibliographic ontology for describing resources	Bibframe
SKOS	Vocabularies and thesauri for knowledge organization	SKOS

Example Ontologies

Ontology	Scope	Link
EpiDoc	Standard for encoding epigraphic texts and inscriptions	EpiDoc
Dublin Core	Metadata standard for describing resources	Dublin Core
Wikidata	Collaborative knowledge base for structured data	Wikidata
OntoUML	Modeling language for formalizing domain knowledge and creating conceptual models	OntoUML
LIDO	Metadata standard for museum objects and cultural heritage	LIDO

Example Ontologies

Ontology	Scope	Link
Kerameikos.org	Standards for intellectual concepts of pottery	Kerameikos
Nomisma.org	Digital representations of numismatic concepts	Nomisma
Linked Places format	Provides a uniform way to build links between different gazetteers	Linked Places

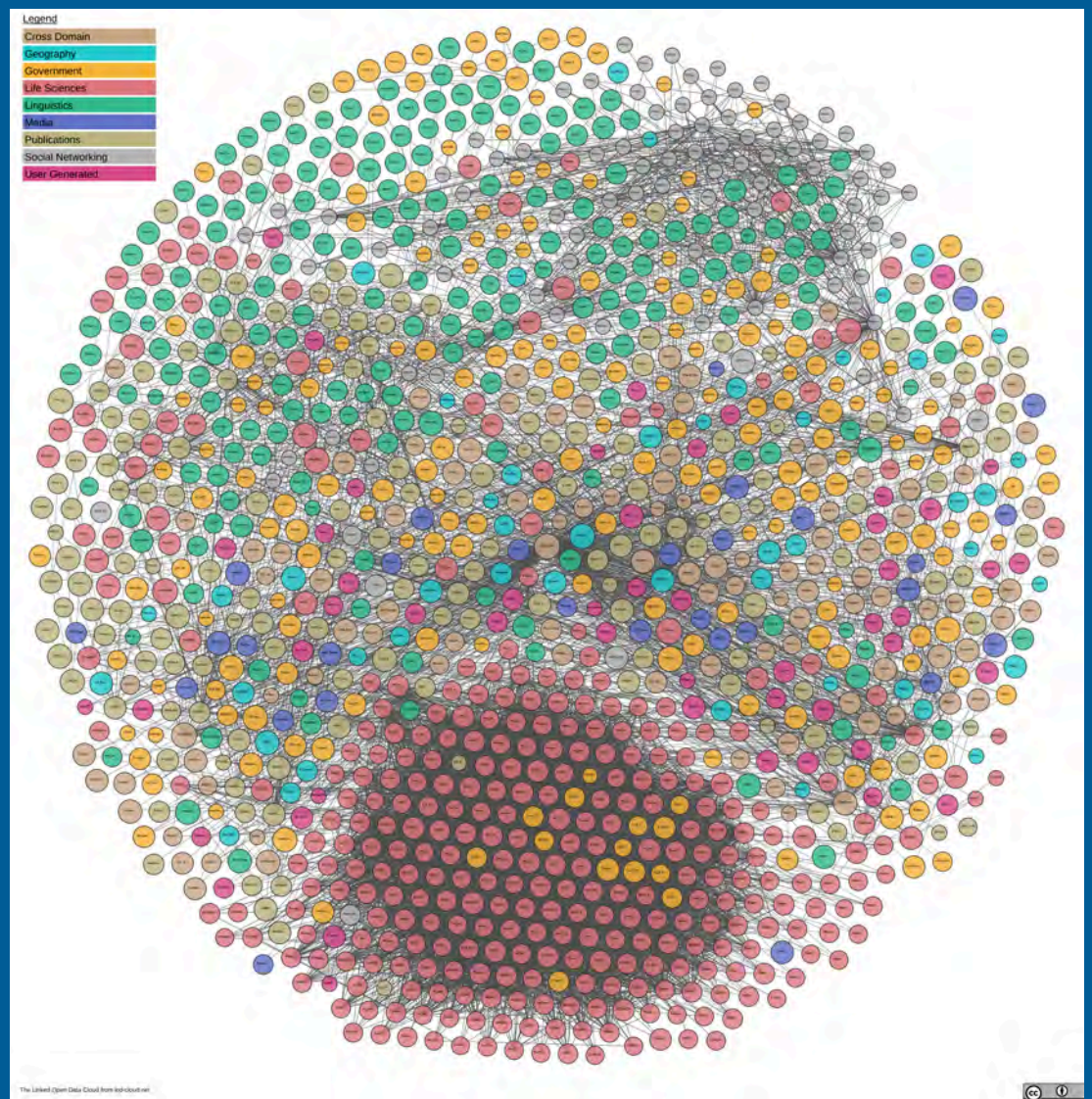
Ontologies

- Ontologies can also be seen as graph based network structures with Classes as the nodes and properties as the edges
- This gets us into the idea of data models

Knowledge Graphs

- When the structure of an ontology is used as the framework to transform data into information, then store and transmit that information
- The application of an ontology to data
- Increasingly tied to the production and consumption of Linked Open Data (LOD)
- Databases are the implementation of knowledge graphs

Linked Open Data (LOD)



LOD

- Text, images, ecology, etc
- Data formats / ontologies from other sources
  - Getty, Library of Congress, and others
- Automatic discovery and communication
  - Anyone using the URIs can discover your data

What is it?

- Essentially: Data on the web that is related to other data
  - Can be explored by a person or computer through links
- 5-Star Data (From Tim Berners-Lee):
  - Available in the web
  - In a machine-readable format
  - Which is non-proprietary
  - Using open standards to identify data, and
  - Linked to other people's / project's data to provide context

RDF, LOD, URIs, and more TLAs!

RDF Sample

```
1 <rdf:Description rdf:ID = “sampleRomanforts/12345”>
2 <rdf:type rdf:resource = “http://www.openannotation.org/ns/Annotati
3 <oac:hasBody rdf:resource = “https://pleiades.stoa.org/places/42302
4 <oac:hasTarget rdf:resource = “http://fake-roman-fort-project/sampl
5 <dcterms:creator rdf:resource = “http://fake-roman-fort-project.org
6 fake-roman-fort-project.org annotation linking http://fake-roman-fo
7 </rdf:Description>
```

RDF

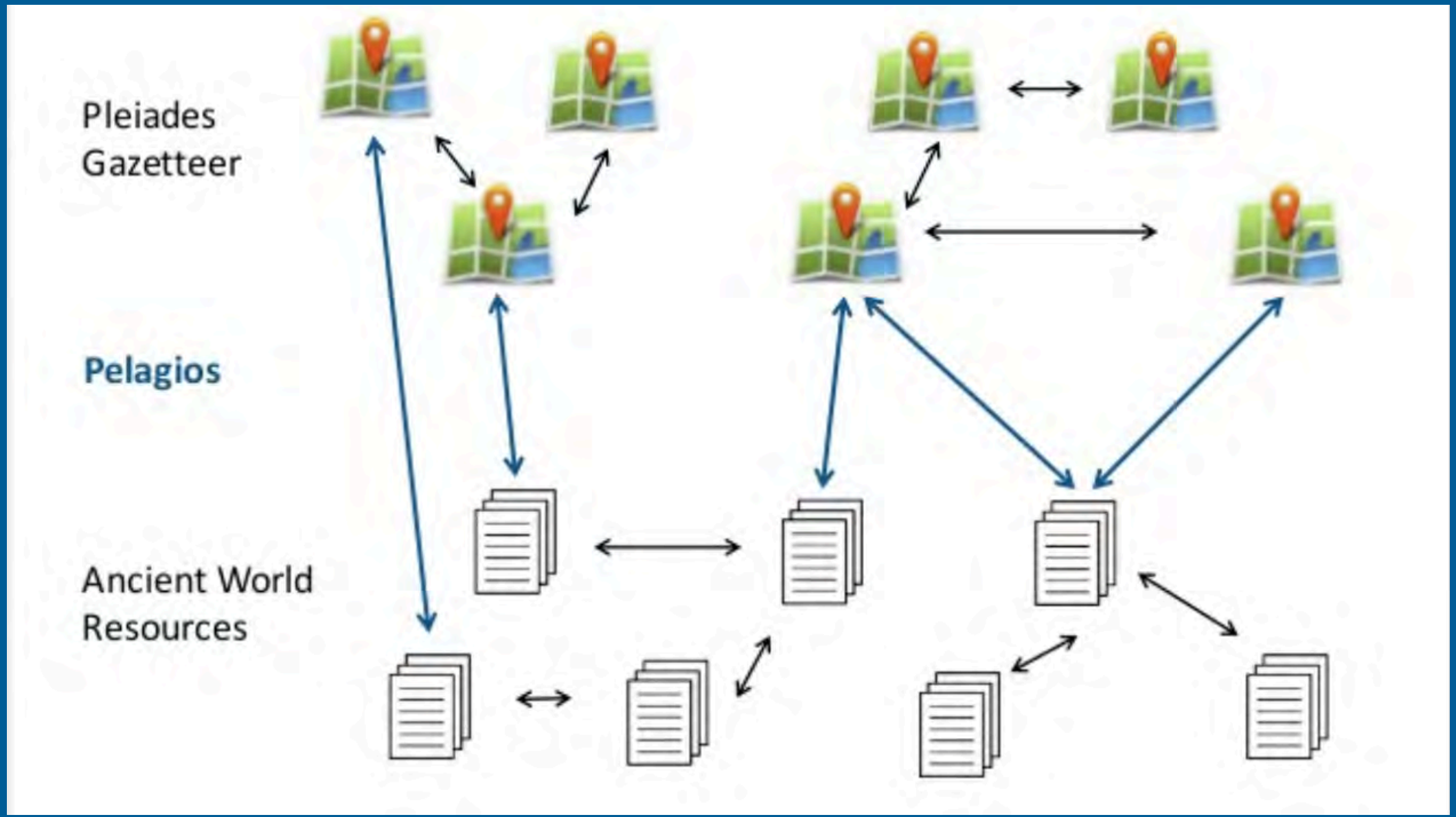
```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
      xmlns:dcterms="http://purl.org/dc/terms/"
      xmlns:void="http://rdfs.org/ns/void#"
      xmlns:foaf="http://xmlns.com/foaf/0.1/"
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<rdf:Description rdf:about="http://pelagios.dme.ait.ac.at/api/datasets/b47df24737cd647789c7d61ede9fed09">
      <rdf:type rdf:resource="http://rdfs.org/ns/void#Dataset"/>
      <dcterms:source rdf:resource="http://arachne.uni-koeln.de/Arachne"/>
      <dcterms:title>Arachne</dcterms:title>
      <dcterms:description>Annotations extracted from Arachne in context of the Pelagios Project.</dcterms:description>
      <dcterms:license rdf:resource="http://creativecommons.org/publicdomain/zero/1.0/"/>
      <void:subset rdf:resource="http://pelagios.dme.ait.ac.at/api/datasets/50f955907f8eb1443de3dac8a28150fb"/>
      <void:subset rdf:resource="http://pelagios.dme.ait.ac.at/api/datasets/b44319c20cb44e11492ee989a9f86b2a"/>
</rdf:Description>
<rdf:Description rdf:about="http://pelagios.dme.ait.ac.at/api/datasets/c4c2ed780a8f4f2b698a5c2505ac34ca">
      <rdf:type rdf:resource="http://rdfs.org/ns/void#Dataset"/>
      <dcterms:source rdf:resource="http://id.clarosnet.org/dataset/claros"/>
      <dcterms:title>CLAROS</dcterms:title>
      <dcterms:description>CLAROS provides searching across a very large virtual database of art and archaeology data, comprising images, texts, antiquarian photographs, and
prosopographic records from Europe, Asia and the East covering 4 millennia.</dcterms:description>
      <dcterms:license rdf:resource="http://opendatacommons.org/licenses/odbl/"/>
      <foaf:homepage rdf:resource="http://www.clarosnet.org/"/>
      <void:subset rdf:resource="http://pelagios.dme.ait.ac.at/api/datasets/7ccf2e63f645de99d7969ad84e5262c7"/>
</rdf:Description>
<rdf:Description rdf:about="http://pelagios.dme.ait.ac.at/api/datasets/21e48d8ca46f666467b81a551fbb1cb">
      <rdf:type rdf:resource="http://rdfs.org/ns/void#Dataset"/>
      <dcterms:source rdf:resource="http://data.perseus.org/annotations/occur/places:pleiades/"/>
      <dcterms:title>Pleiades Annotations in the Perseus Digital Library</dcterms:title>
      <dcterms:description>Occurrences of places identified by Pleiades URIs in texts from the Perseus Digital Library</dcterms:description>
      <dcterms:license rdf:resource="http://creativecommons.org/licenses/by-sa/3.0/us/"/>
      <void:subset rdf:resource="http://pelagios.dme.ait.ac.at/api/datasets/cf3baed0f33f28c02dbed50a96b663af"/>
</rdf:Description>
```

URI

- URI (Uniform Resource Identifier)
- Used to identify an entity in a hierarchical system
- Familiar to us with URLs

Pelagios and the LAWDI community

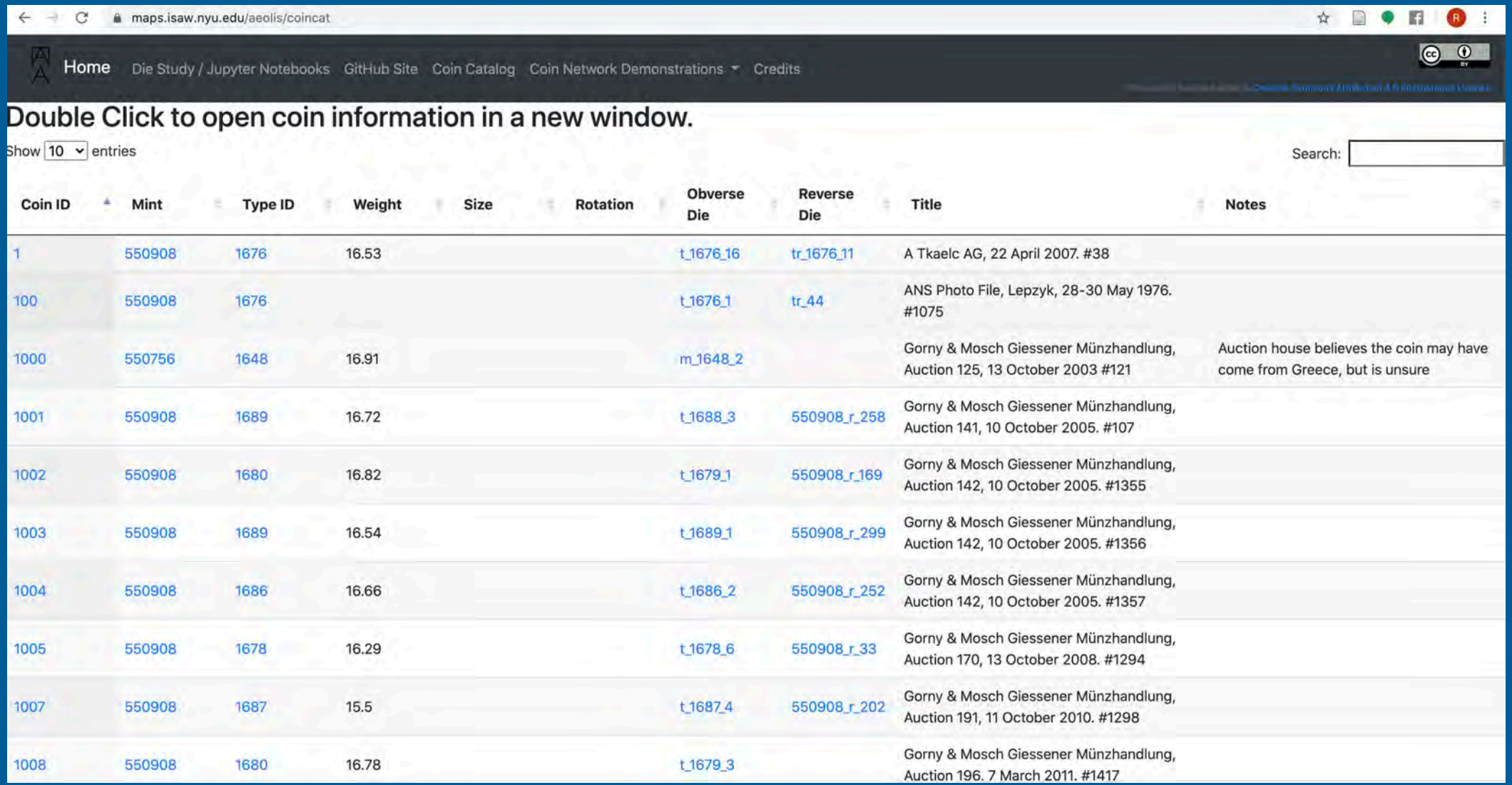
- Pelagios
- http://commons.pelagios.org/
- Explores / Links data on historical places
- LAWDI (Linked Ancient World Data Institute / Initiative)
  - #lawdi on Twitter
  - From a NEH grant – community and practice



Example: Networks of Coins

- Data sources
  - Pleiades = spatial data for cities
  - American Numismatic Society (ANS): MANTIS, NOMISMA, Coins, spatial information, data structures, some hoard data
  - Museums using ANS data structures
  - Ancient World Mapping Center = spatial data (roads, etc)

id	hoard_id	title	ex_start_date	ex_end_date	buried	b_start_date	b_end_date	contents	notes	location_uri
1	ch1,1	Asia Minor	1974		-550--545	-550.0	-545.0	6+ st., heavy standard		https://pleiades.stoa.org/places/837
2	ch1,2	Hellespont	1969		c. 550 B.C.?	-550.0	-550.0	49+ EL		https://pleiades.stoa.org/places/501434
3	ch1,3	Asia Minor	1935/1940		c. 525 B.C.	-525.0	-525.0	906 AR and hacksilber		https://pleiades.stoa.org/places/837
4	ch1,4	South Anatolia	1971		c. 500 B.C.	-500.0	-500.0	10+ AR		https://www.geonames.org/10922502
5	ch1,5	Aegina	1973		c. 500 B.C.?	-500.0	-500.0	12+ AR		https://www.geonames.org/265501
...	...	...	...	...	...	...	...	...	...	...
2365	ch10,247	Unknown findspot	1993			NaN	NaN	42 AU		
4843	igch2254	Cibali, Catana district	1907			NaN	NaN	10 AE		https://www.geonames.org/2524993
4845	igch2256	Bronte	1906			NaN	NaN	1+ EL		https://www.geonames.org/2525498
4846	igch2257	Lipari	1920			NaN	NaN	6 AE		https://www.geonames.org/2524378
4847	igch2258	Termini Imerese environs	1958			NaN	NaN	16 AR		https://www.geonames.org/2522960


maps.isaw.nyu.edu/aeolis/coincat
Home
Die Study / Jupyter Notebooks
GitHub Site
Coin Catalog
Coin Network Demonstrations
Credits
Double Click to open coin information in a new window.
Show 10 entries
Search:
Coin ID
Mint
Type ID
Weight
Size
Rotation
Obverse Die
Reverse Die
Title
Notes
1 550908 1676 16.53 t_1676_16 tr_1676_11 A Tkaelc AG, 22 April 2007. #38
100 550908 1676 t_1676_1 tr_44 ANS Photo File, Lepzyk, 28-30 May 1976. #1075
1000 550756 1648 16.91 m_1648_2 Gorny & Mosch Giessener Münzhandlung, Auction 125, 13 October 2003 #121 Auction house believes the coin may have come from Greece, but is unsure
1001 550908 1689 16.72 t_1688_3 550908_r_258 Gorny & Mosch Giessener Münzhandlung, Auction 141, 10 October 2005. #107
1002 550908 1680 16.82 t_1679_1 550908_r_169 Gorny & Mosch Giessener Münzhandlung, Auction 142, 10 October 2005. #1355
1003 550908 1689 16.54 t_1689_1 550908_r_299 Gorny & Mosch Giessener Münzhandlung, Auction 142, 10 October 2005. #1356
1004 550908 1686 16.66 t_1686_2 550908_r_252 Gorny & Mosch Giessener Münzhandlung, Auction 142, 10 October 2005. #1357
1005 550908 1678 16.29 t_1678_6 550908_r_33 Gorny & Mosch Giessener Münzhandlung, Auction 170, 13 October 2008. #1294
1007 550908 1687 15.5 t_1687_4 550908_r_202 Gorny & Mosch Giessener Münzhandlung, Auction 191, 11 October 2010. #1298
1008 550908 1680 16.78 t_1679_3 Gorny & Mosch Giessener Münzhandlung, Auction 196, 7 March 2011. #1417

jupyter aa_writeup Last Checkpoint: a day ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

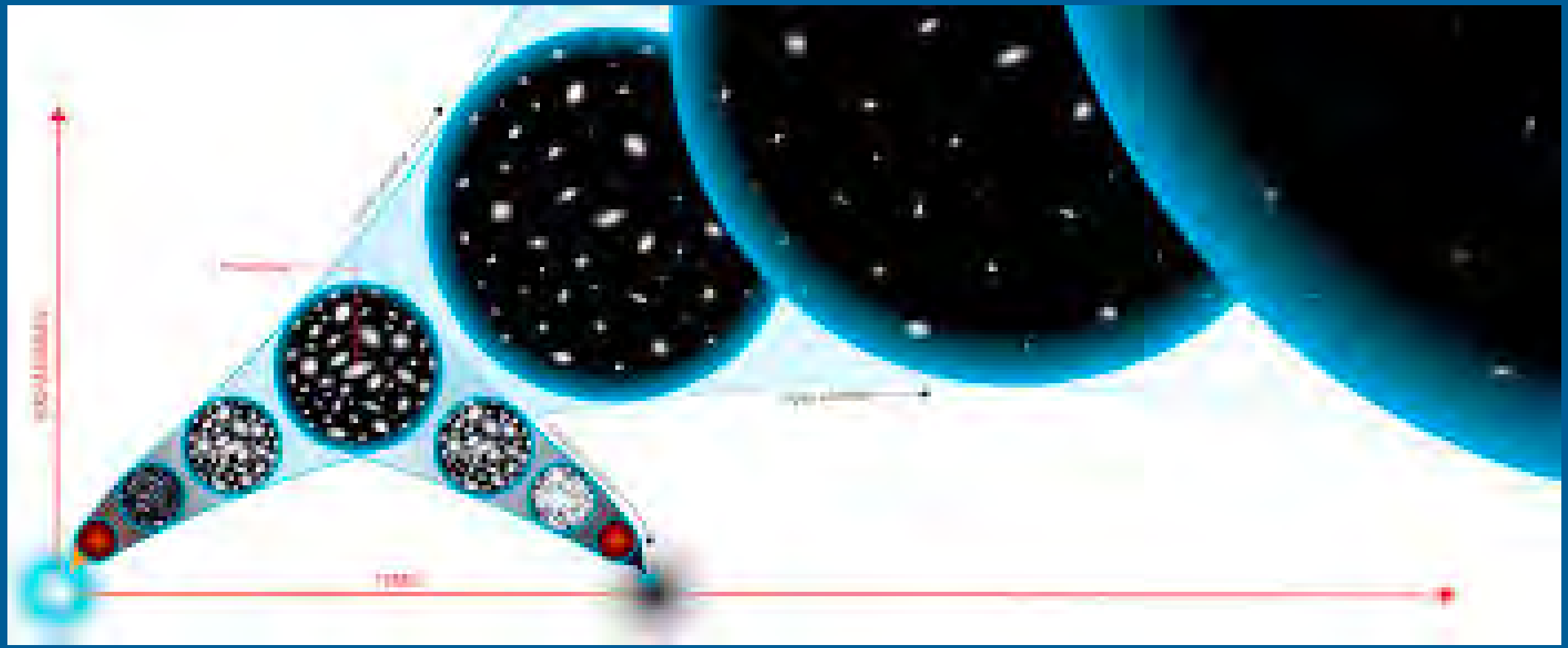
Run Markdown Cite

```
addNetworkRoutesFromFile(temnos_routes['url'], temnos_routes['featuregroupname'], temnos_routes['sizeproperty'], 'black
addNetworkRoutesFromFile(myrina_routes['url'], myrina_routes['featuregroupname'], myrina_routes['sizeproperty'], 'pink'
addNetworkRoutesFromFile(kyme_routes['url'], kyme_routes['featuregroupname'], kyme_routes['sizeproperty'], 'purple', 20



addFeatureGrouptoMap(temnoshoardsdf, hoardsfg, 30, 5, 'degree', 'title', m)
m.add_child(folium.map.LayerControl())

m
```

Out[33]:

The extensive Temnian minting during the time, especially the very large production of type 1690, could have been an extremely beneficial resource for the Attalids to draw upon in support of their own political projects. One attractive possibility for Temnian coins was their use as a medium to pay mercenaries, as soldiers for hire would presumably want some form of wealth that was transferable outside of the Attalid kingdom *(Le Rider 1989)*. Such a consideration is clearly shown in an earlier mutiny of mercenaries at Philateria and Attalea against the Attalid monarch Eumenes I (r. 263-241) (*OGIS* 226). As part of a compact ending the disturbance, Eumenes swore to provide his rioting troops with fixed prices for food, back pay, and the ability for a discharged soldier to "take his

How do I do this before the heat death of the universe?



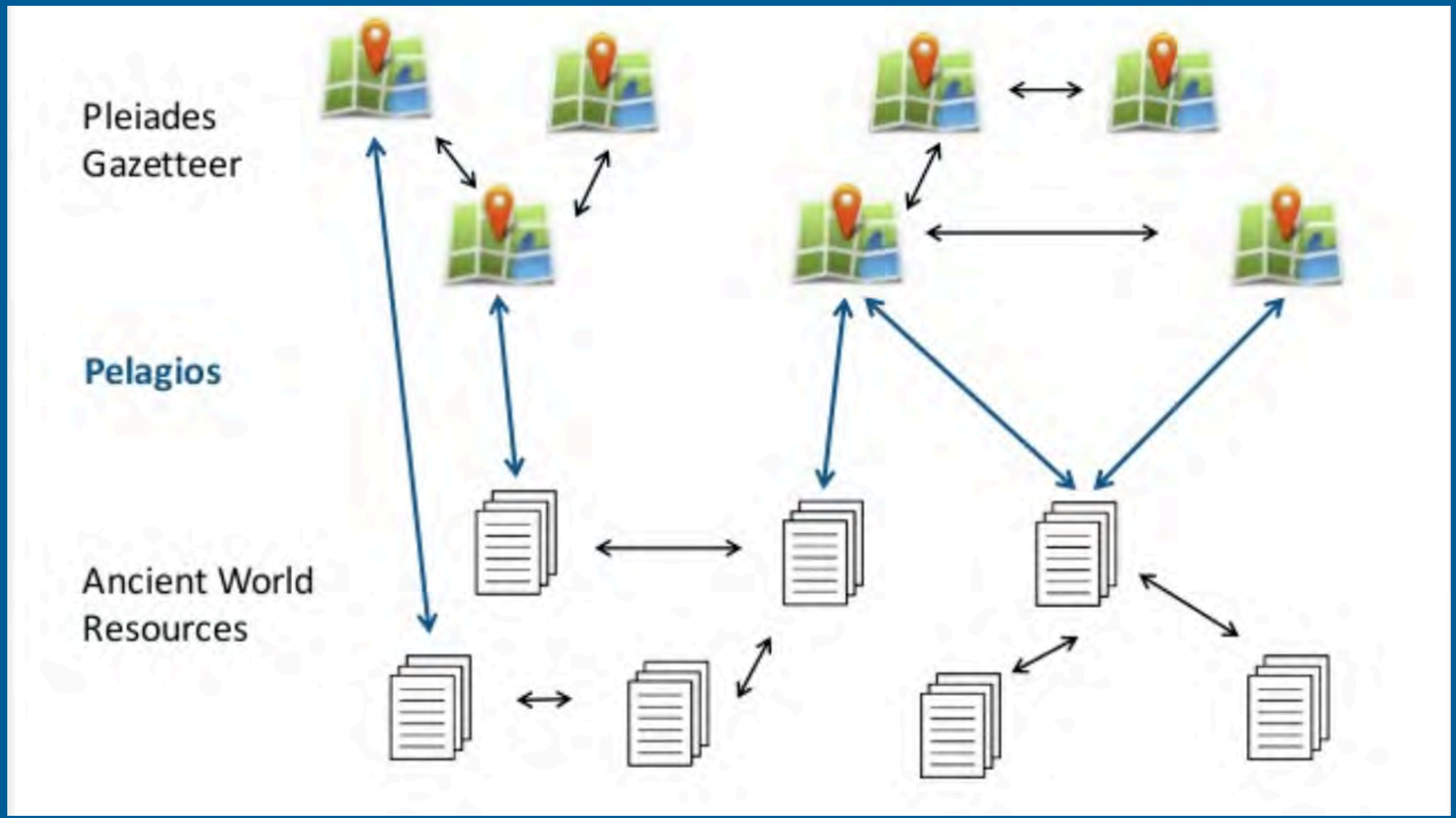
Reconcile Your Data!

- Your own ids can be associated with information from other LOD projects like digital gazetteers, prosopographies, etc
- You then get all of the relevant information
  - Along with other info!
- Can be done by hand
  - Or with tools like Recogito
- Think about *what* you are reconciling to
  - Wikidata is a fantastic choice!

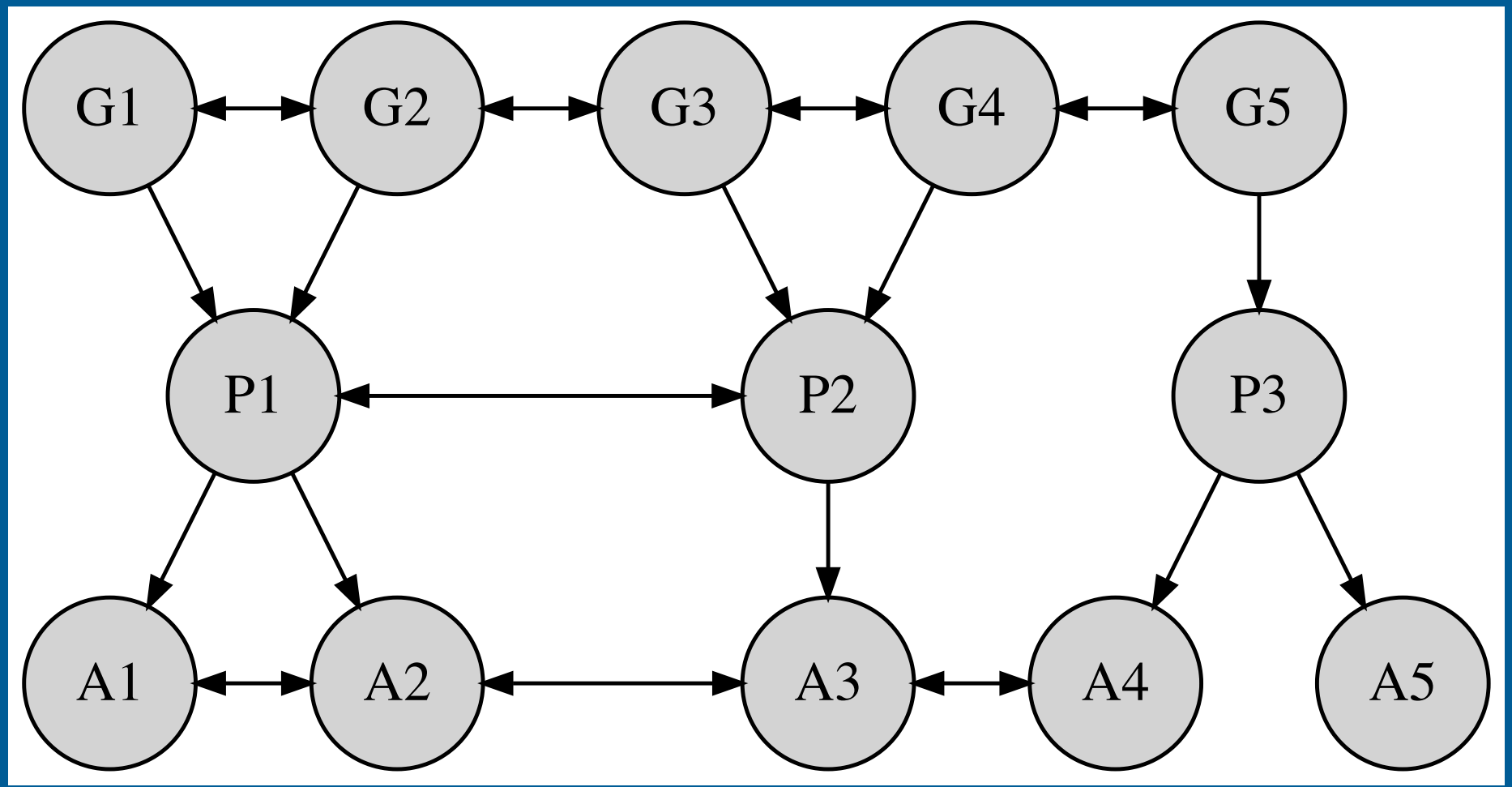
Example Gazetteers

Gazetteer	Scope	Link
Pleiades	Geospatial ontology for ancient places and geographical features	Pleiades
PeriodO	A public domain gazetteer of scholarly definitions of historical, art-historical, and archaeological periods.	PeriodO
World History Gazetteer (WHG)	A platform for linking records about historical places	WHG
Syriaca.org: The Syriac Reference Portal	A a reference hub for digitally linking research findings	Syriaca.org

Return To the Pelagios *Network*



Return To the Pelagios *Network*



Network-Centric Approach

- Choice to model information as connected nodes in a graph
- Assumes that the understanding and analysis of these entities cannot be untangled from their relationships with one another
- Networks can be spatial, incorporate temporal information, and contain any number of conceptually different entities

Basic Idea

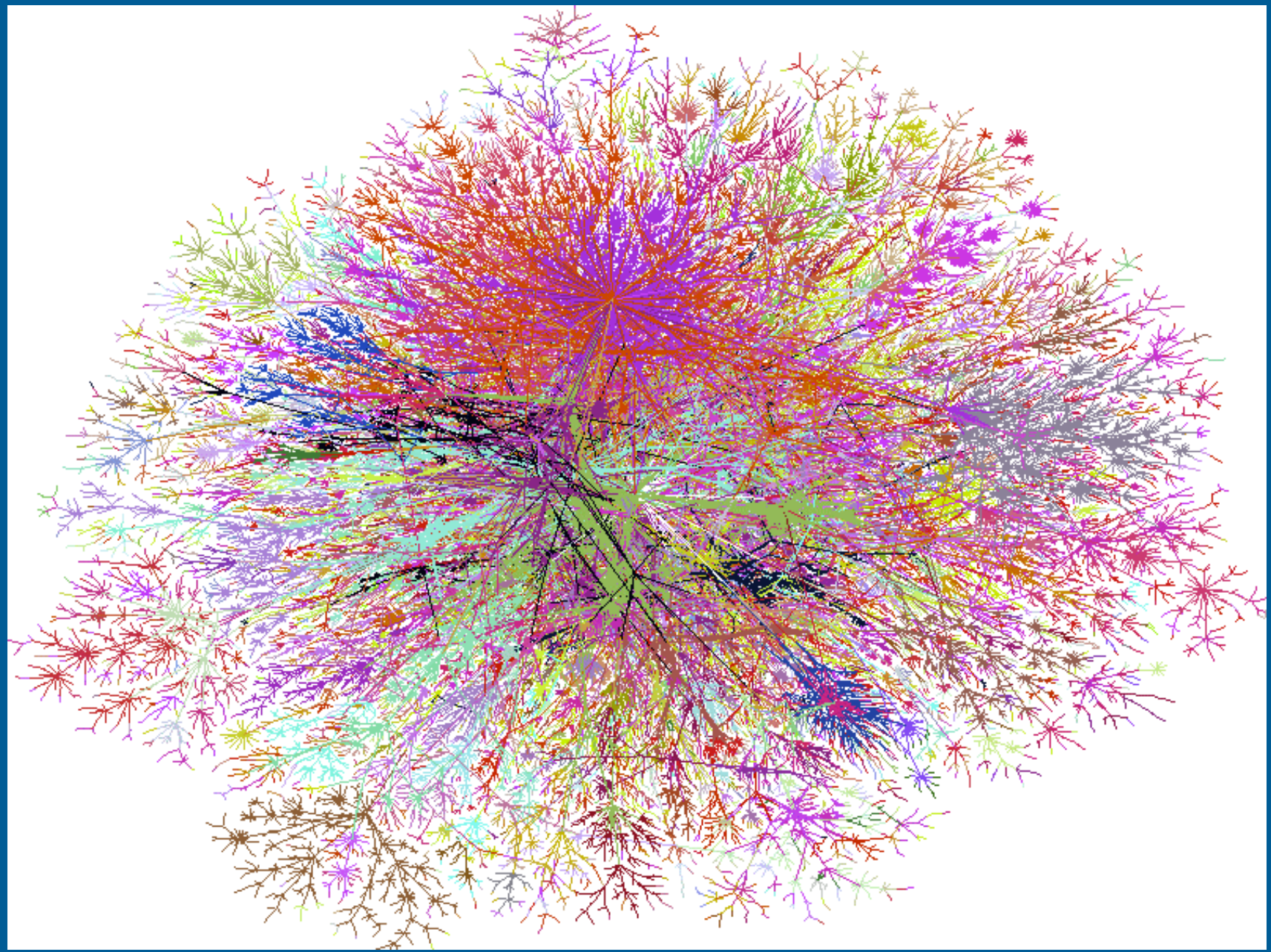
- Networks are a collection of entities
- At least some are *linked*
- All kinds of subject domains
- Very flexible definition

What is SNA?

- Method to perform visual and mathematical analysis of relationships
- Analysis has to be visually interesting / useful and mathematically rigorous
- Fundamental point: We are looking at networks
- Connections, connections, connections
- Networks are all over the place

What is SNA Used For?

- People
- Concepts
- Literature
- Biological Systems
- Electronic Systems



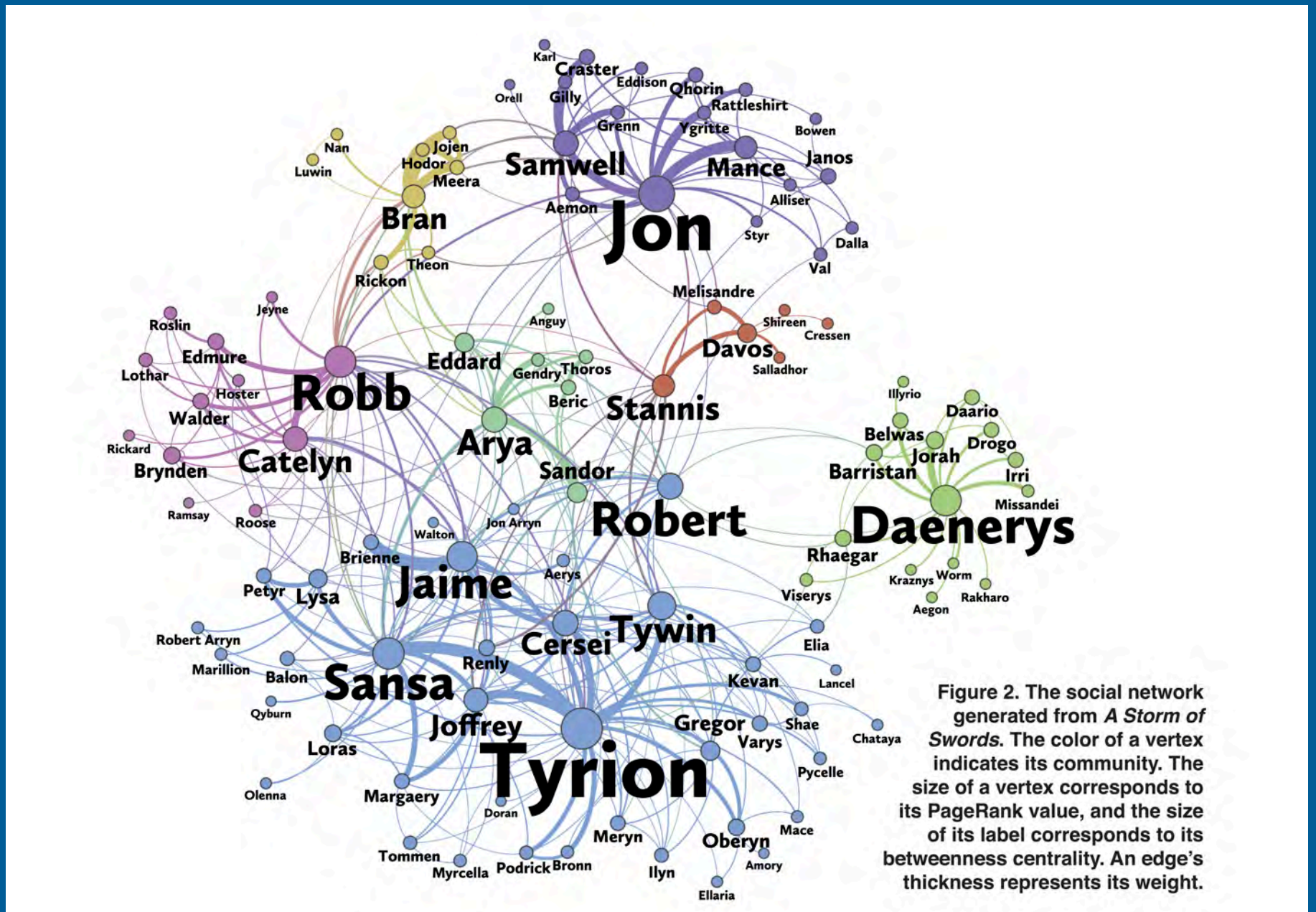
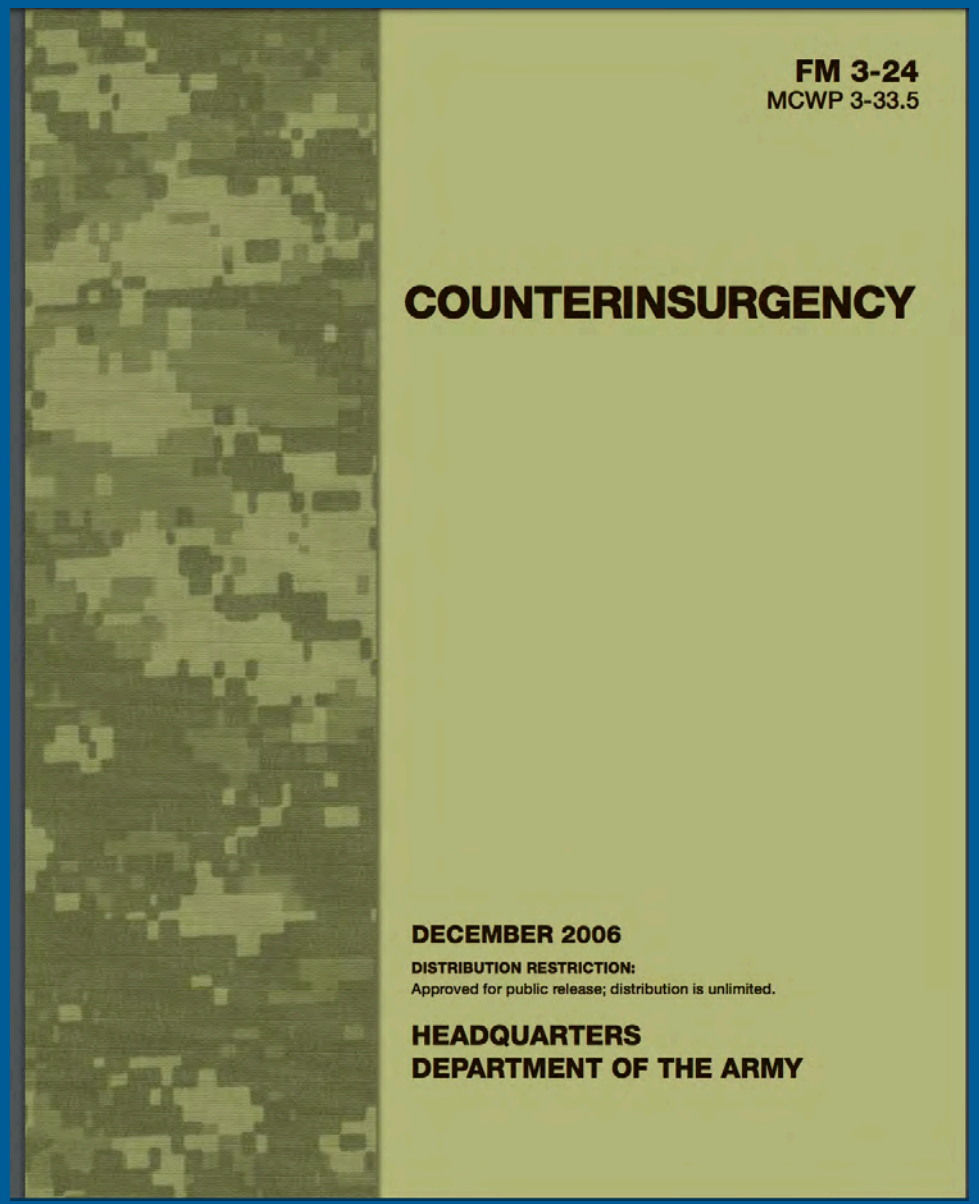


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.



Events

B-26. Events are routine, cyclical, planned, or spontaneous activities that significantly affect organizations, people, and military operations. They are often symbols, as described in paragraph 3-51. Examples include the following:

- National and religious holidays.
- Agricultural crop, livestock, and market cycles.
- Elections.
- Civil disturbances.
- Celebrations.

B-27. Other events include disasters from natural, manmade, or technological sources. These create hardships and require emergency responses. Examples of events precipitated by military forces include combat operations, deployments, redeployments, and paydays. Once significant events are determined, it is important to template the events and analyze them for their political, economic, psychological, environmental, and legal implications.

EVALUATE THE THREAT

B-28. Evaluating the threat involves analyzing insurgent organizations, capabilities, and tactics to identify vulnerabilities to exploit. Tools like social network analysis, link diagrams, and association matrices help analysts do this. Other tools such as historical time lines and pattern analysis tools assist in developing event and doctrinal templates to depict enemy tactics.

Social Network Analysis

B-29. Social network analysis (SNA) is a tool for understanding the organizational dynamics of an insurgency and how best to attack or exploit it. It allows analysts to identify and portray the details of a network structure. Its shows how an insurgency's networked organization behaves and how that connectivity affects its behavior. SNA allows analysts to assess the network's design, how its member may or may not act autonomously, where the leadership resides or how it is distributed among members, and how hierarchical dynamics may mix or not mix with network dynamics.

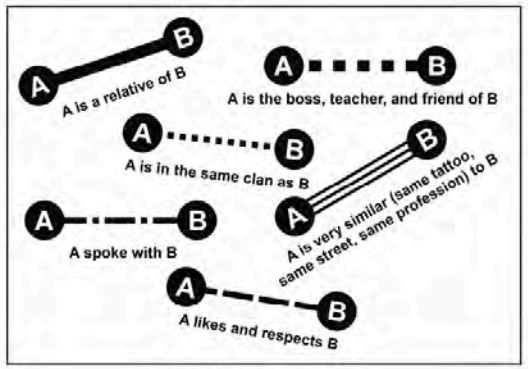


Figure B-5. Examples of dyads

Basic Terms

Nodes

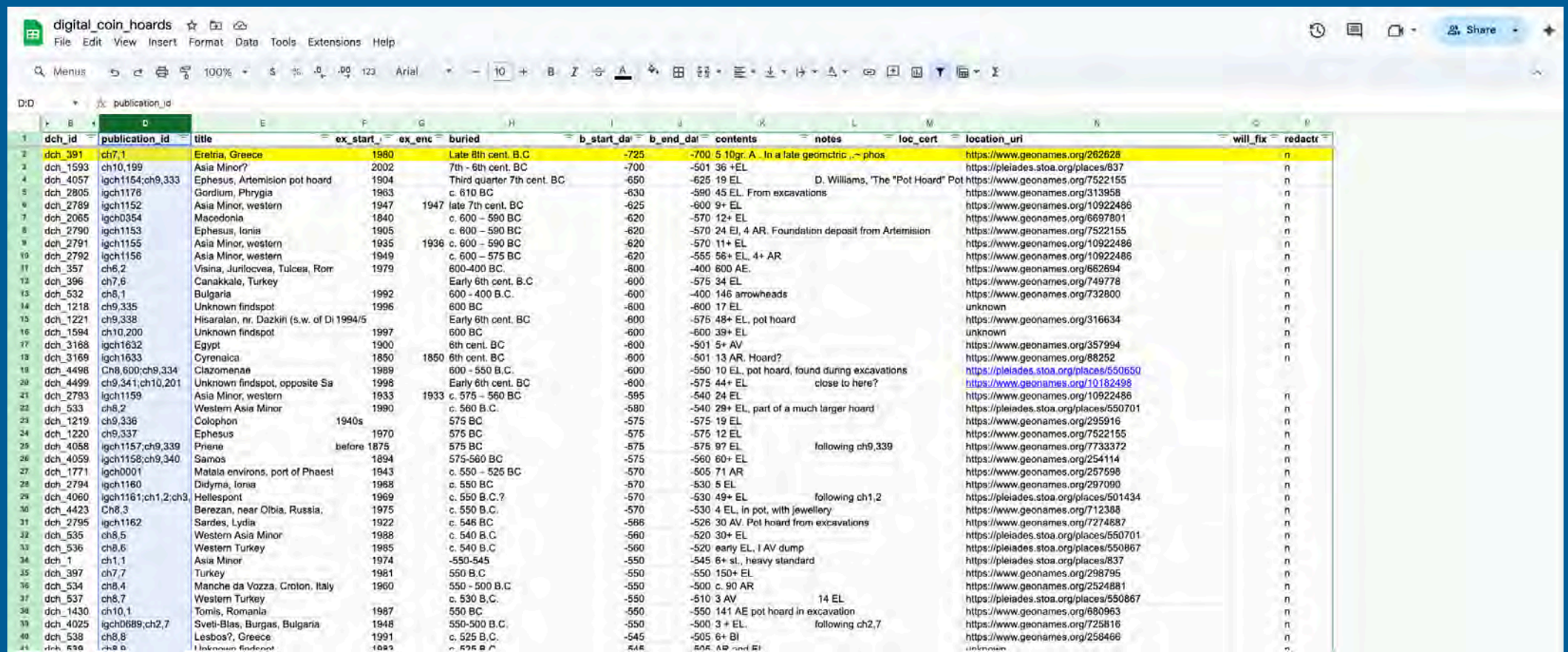
Edges

Graph

Nodes

- The “stuff” you are looking at
- Any attributes you want!
- Traditionally represented by circles on a graph
- May see “actors” used

Nodes



digital_coin_hoards

File Edit View Insert Format Data Tools Extensions Help

Share

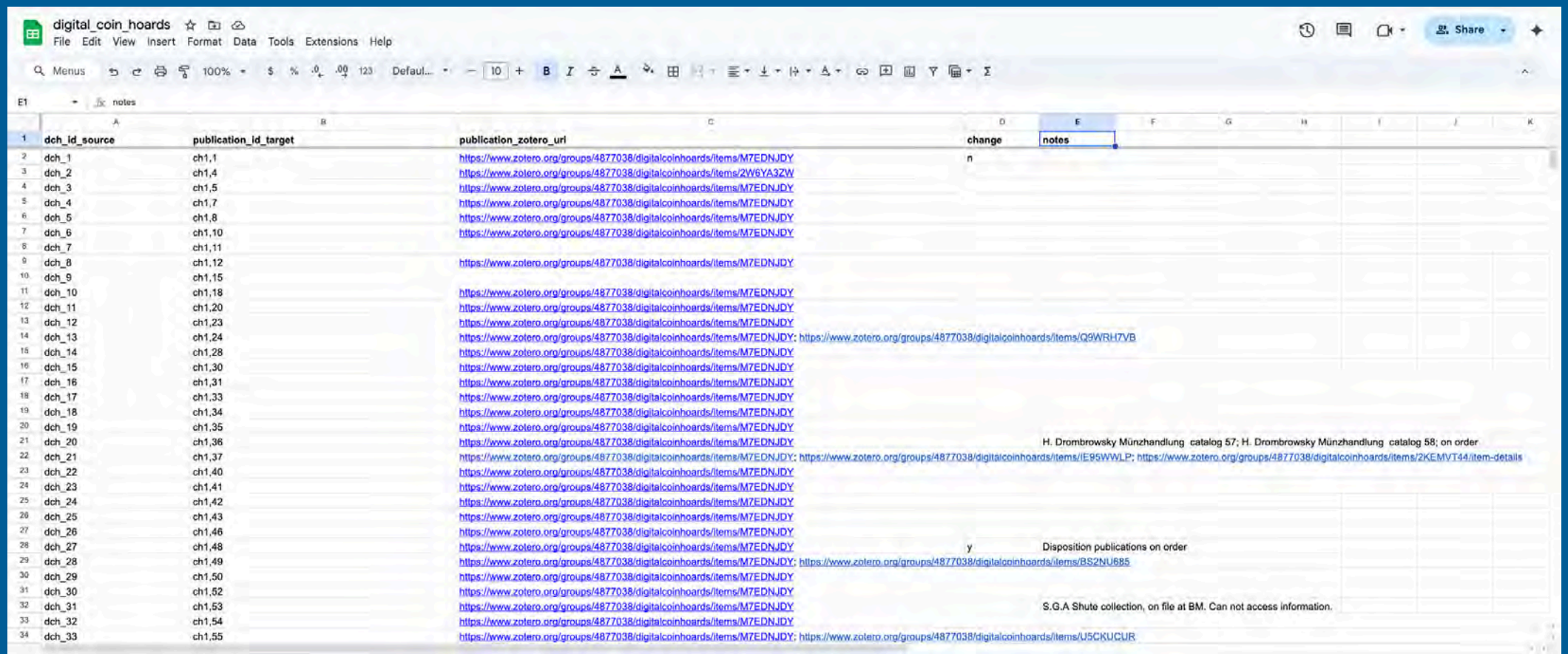
D:D publication_id

dch_id	publication_id	title	ex_start_	ex_enc	buried	b_start_da	b_end_dat	contents	notes	loc_cert	location_uri	will_fix	redacte
dch_391	ch7,1	Eretria, Greece	1980		Late 8th cent. B.C	-725	-700	5 10gr. A . In a late geometric ,.~ phos			https://www.geonames.org/262628		n
dch_1593	ch10,199	Asia Minor?	2002		7th - 6th cent. BC	-700	-501	36 +EL			https://pleiades.stoa.org/places/837		n
dch_4057	igch1154;ch9,333	Ephesus, Artemision pot hoard	1904		Third quarter 7th cent. BC	-650	-625	19 EL	D. Williams, "The "Pot Hoard" Pot		https://www.geonames.org/7522155		n
dch_2805	igch1176	Gordium, Phrygia	1963		c. 610 BC	-630	-590	45 EL. From excavations			https://www.geonames.org/313958		n
dch_2789	igch1152	Asia Minor, western	1947	1947	late 7th cent. BC	-625	-600	9+ EL			https://www.geonames.org/10922486		n
dch_2065	igch0354	Macedonia	1840		c. 600 – 590 BC	-620	-570	12+ EL			https://www.geonames.org/6697801		n
dch_2790	igch1153	Ephesus, Ionia	1905		c. 600 – 590 BC	-620	-570	24 El, 4 AR. Foundation deposit from Artemision			https://www.geonames.org/7522155		n
dch_2791	igch1155	Asia Minor, western	1935	1936	c. 600 – 590 BC	-620	-570	11+ EL			https://www.geonames.org/10922486		n
dch_2792	igch1156	Asia Minor, western	1949		c. 600 – 575 BC	-620	-555	56+ EL. 4+ AR			https://www.geonames.org/10922486		n
dch_357	ch6,2	Visina, Jurilocvea, Tulcea, Rom	1979		600-400 BC.	-600	-400	600 AE.			https://www.geonames.org/662694		n
dch_396	ch7,6	Canakkale, Turkey			Early 6th cent. B.C	-600	-575	34 EL			https://www.geonames.org/749778		n
dch_532	ch8,1	Bulgaria	1992		600 - 400 B.C.	-600	-400	146 arrowheads			https://www.geonames.org/732800		n
dch_1218	ch9,335	Unknown findspot	1996		600 BC	-600	-600	17 EL			unknown		n
dch_1221	ch9,338	Hisaralan, nr. Dazkiri (s.w. of Di	1994/5		Early 6th cent. BC	-600	-575	48+ EL, pot hoard			https://www.geonames.org/316634		n
dch_1594	ch10,200	Unknown findspot	1997		600 BC	-600	-600	39+ EL			unknown		n
dch_3168	igch1632	Egypt	1900		6th cent. BC	-600	-501	5+ AV			https://www.geonames.org/357994		n
dch_3169	igch1633	Cyrenaica	1850	1850	6th cent. BC	-600	-501	13 AR. Hoard?			https://www.geonames.org/88252		n
dch_4498	Ch8,600;ch9,334	Clazomenae	1989		600 - 550 B.C.	-600	-550	10 EL, pot hoard, found during excavations			https://pleiades.stoa.org/places/550650		
dch_4499	ch9,341;ch10,201	Unknown findspot, opposite Sa	1998		Early 6th cent. BC	-600	-575	44+ EL	close to here?		https://www.geonames.org/10182498		
dch_2793	igch1159	Asia Minor, western	1933	1933	c. 575 – 560 BC	-595	-540	24 EL			https://www.geonames.org/10922486		n
dch_533	ch8,2	Western Asia Minor	1990		c. 560 B.C.	-580	-540	29+ EL, part of a much larger hoard			https://pleiades.stoa.org/places/550701		n
dch_1219	ch9,336	Colophon	1940s		575 BC	-575	-575	19 EL			https://www.geonames.org/295916		n
dch_1220	ch9,337	Ephesus	1970		575 BC	-575	-575	12 EL			https://www.geonames.org/7522155		n
dch_4058	igch1157;ch9,339	Priene	before 1875		575 BC	-575	-575	97 EL	following ch9,339		https://www.geonames.org/7733372		n
dch_4059	igch1158;ch9,340	Samos	1894		575-560 BC	-575	-560	60+ EL			https://www.geonames.org/254114		n
dch_1771	igch0001	Matala environs, port of Phaest	1943		c. 550 – 525 BC	-570	-505	71 AR			https://www.geonames.org/257598		n
dch_2794	igch1160	Didyma, Ionia	1968		c. 550 BC	-570	-530	5 EL			https://www.geonames.org/297090		n
dch_4060	igch1161;ch1,2;ch3,	Hellespont	1969		c. 550 B.C.?	-570	-530	49+ EL	following ch1,2		https://pleiades.stoa.org/places/501434		n
dch_4423	Ch8,3	Berezan, near Olbia, Russia,	1975		c. 550 B.C.	-570	-530	4 EL, in pot, with jewellery			https://www.geonames.org/712388		n
dch_2795	igch1162	Sardes, Lydia	1922		c. 546 BC	-566	-526	30 AV. Pot hoard from excavations			https://www.geonames.org/7274687		n
dch_535	ch8,5	Western Asia Minor	1988		c. 540 B.C	-560	-520	30+ EL			https://pleiades.stoa.org/places/550701		n
dch_536	ch8,6	Western Turkey	1985		c. 540 B.C	-560	-520	early EL, I AV dump			https://pleiades.stoa.org/places/550867		n
dch_1	ch1,1	Asia Minor	1974		-550-545	-550	-545	6+ sl., heavy standard			https://pleiades.stoa.org/places/837		n
dch_397	ch7,7	Turkey	1981		550 B.C	-550	-550	150+ EL			https://www.geonames.org/298795		n
dch_534	ch8,4	Manche da Vozza, Croton, Italy	1960		550 - 500 B.C	-550	-500	c. 90 AR			https://www.geonames.org/2524881		n
dch_537	ch8,7	Western Turkey			c. 530 B.C.	-550	-510	3 AV	14 EL		https://pleiades.stoa.org/places/550867		n
dch_1430	ch10,1	Tomis, Romania	1987		550 BC	-550	-550	141 AE pot hoard in excavation			https://www.geonames.org/680963		n
dch_4025	igch0689;ch2,7	Sveti-Blas, Burgas, Bulgaria	1948		550-500 B.C.	-550	-500	3 + EL.	following ch2,7		https://www.geonames.org/725816		n
dch_538	ch8,8	Lesbos?, Greece	1991		c. 525 B.C.	-545	-505	6+ BI			https://www.geonames.org/258466		n
dch_539	ch8,9	Unknown findspot	1983		c. 525 B.C.	-545	-505	AR and EL			unknown		n

Edges

- What links your *nodes*
  - Sometimes called links
  - Up to us to define what this means
- Think of this workshop
- Can have *weight*
- Lines on a *graph*

Edges



digital_coin_hoards

File Edit View Insert Format Data Tools Extensions Help

Share

E1 notes

	A	B	C	D	E	F	G	H	I	J	K
1	dch_id_source	publication_id_target	publication_zotero_url	change	notes						
2	dch_1	ch1,1	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY	n							
3	dch_2	ch1,4	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/2W6YA3ZW								
4	dch_3	ch1,5	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
5	dch_4	ch1,7	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
6	dch_5	ch1,8	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
7	dch_6	ch1,10	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
8	dch_7	ch1,11									
9	dch_8	ch1,12	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
10	dch_9	ch1,15									
11	dch_10	ch1,18	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
12	dch_11	ch1,20	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
13	dch_12	ch1,23	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
14	dch_13	ch1,24	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY; https://www.zotero.org/groups/4877038/digitalcoinhoards/items/Q9WRH7VB								
15	dch_14	ch1,28	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
16	dch_15	ch1,30	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
17	dch_16	ch1,31	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
18	dch_17	ch1,33	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
19	dch_18	ch1,34	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
20	dch_19	ch1,35	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
21	dch_20	ch1,36	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY		H. Drombrowsky Münzhandlung catalog 57; H. Drombrowsky Münzhandlung catalog 58; on order						
22	dch_21	ch1,37	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY; https://www.zotero.org/groups/4877038/digitalcoinhoards/items/IE95WWLP; https://www.zotero.org/groups/4877038/digitalcoinhoards/items/2KEMVT44/item-details								
23	dch_22	ch1,40	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
24	dch_23	ch1,41	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
25	dch_24	ch1,42	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
26	dch_25	ch1,43	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
27	dch_26	ch1,46	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
28	dch_27	ch1,48	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY	y	Disposition publications on order						
29	dch_28	ch1,49	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY; https://www.zotero.org/groups/4877038/digitalcoinhoards/items/BS2NU685								
30	dch_29	ch1,50	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
31	dch_30	ch1,52	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
32	dch_31	ch1,53	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY		S.G.A Shute collection, on file at BM. Can not access information.						
33	dch_32	ch1,54	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY								
34	dch_33	ch1,55	https://www.zotero.org/groups/4877038/digitalcoinhoards/items/M7EDNJDY; https://www.zotero.org/groups/4877038/digitalcoinhoards/items/U5CKUCUR								

Graph

- Mathematical models of network structures
- There are rules!

Undirected Graph

- Direction of the relationship does not matter
- No arrows are needed
- “Default” for many SNA discussions

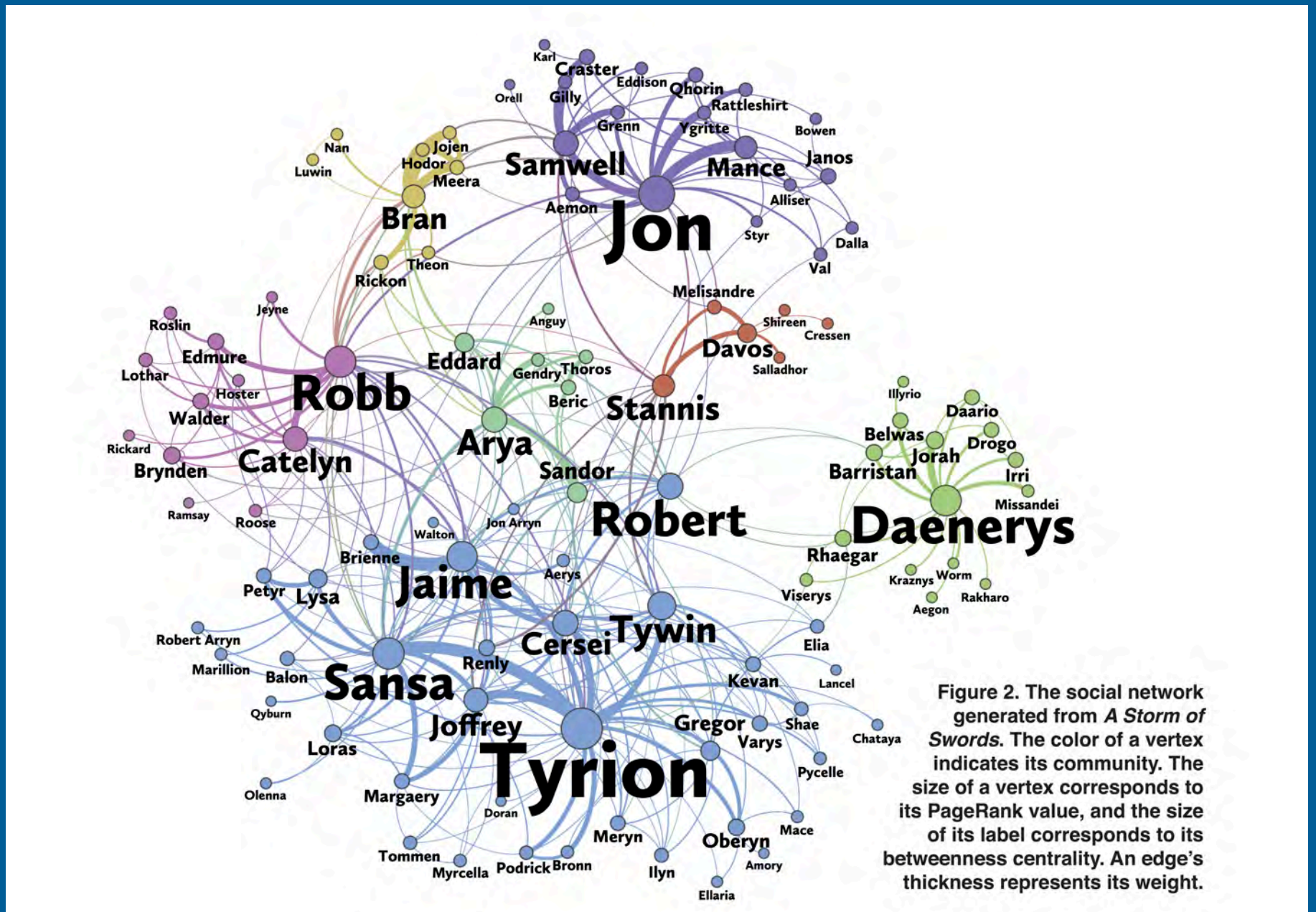
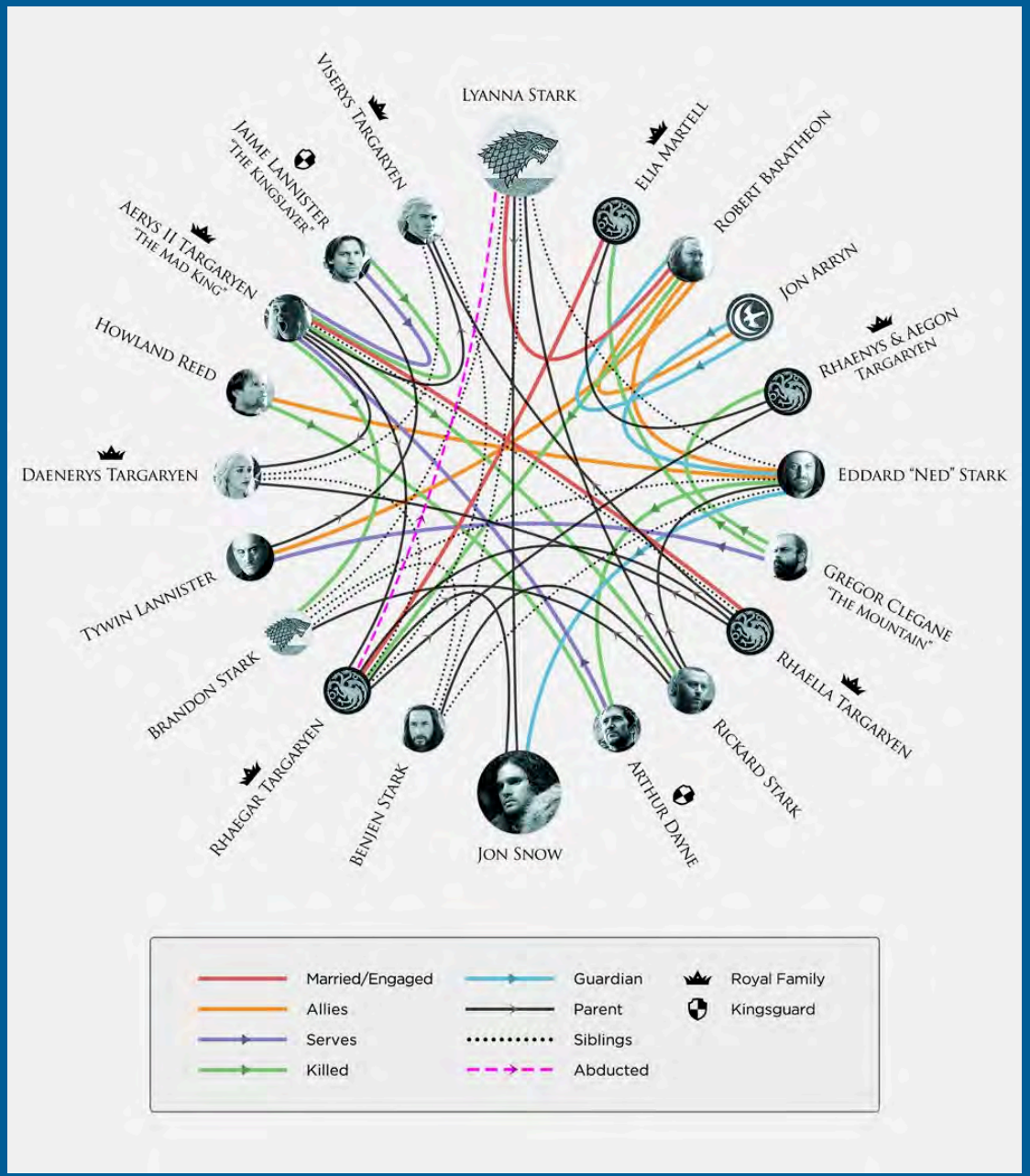


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.

Symmetrical and Asymmetrical Edges

- Sometimes not enough to simply show a connection
- A → B but not the other way around
- Directed graph: Directed nodes and directed edges


Lyanna Stark
Viserys Targaryen
Jaime Lannister "The Kingslayer"
Aerys II Targaryen "The Mad King"
Howland Reed
Daenerys Targaryen
Tywin Lannister
Brandon Stark
Rhaegar Targaryen
Benjen Stark
Jon Snow
Arthur Dayne
Rickard Stark
Rhaella Targaryen
Gregor Clegane "The Mountain"
Eddard "Ned" Stark
Rhaenys & Aegon Targaryen
Jon Arryn
Robert Baratheon
Elia Martell
Married/Engaged
Allies
Serves
Killed
Guardian
Parent
Siblings
Abducted
Royal Family
Kingsguard

Modes vs Types

- Modes tell you what is in the graph
  - People, people and things, etc
- Types tell you what kind of graph it is
  - Are we looking at an individual? Group? Text?

One Mode / Unipartite Networks

- Relationships among a single set of similar nodes
- Same type of nodes
- Same type of edges
  - Some differentiation allowed; i.e. parent / child relationships

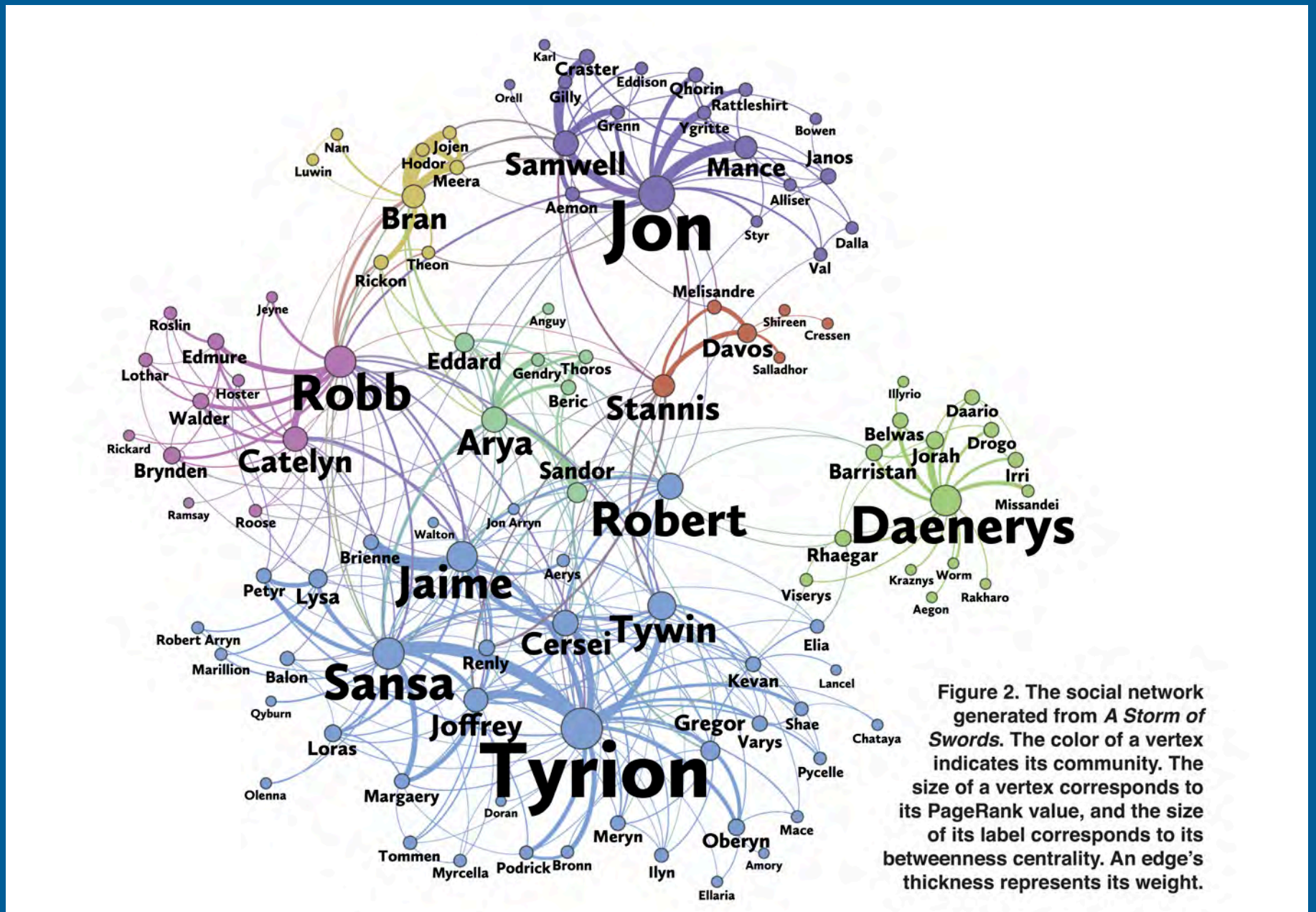
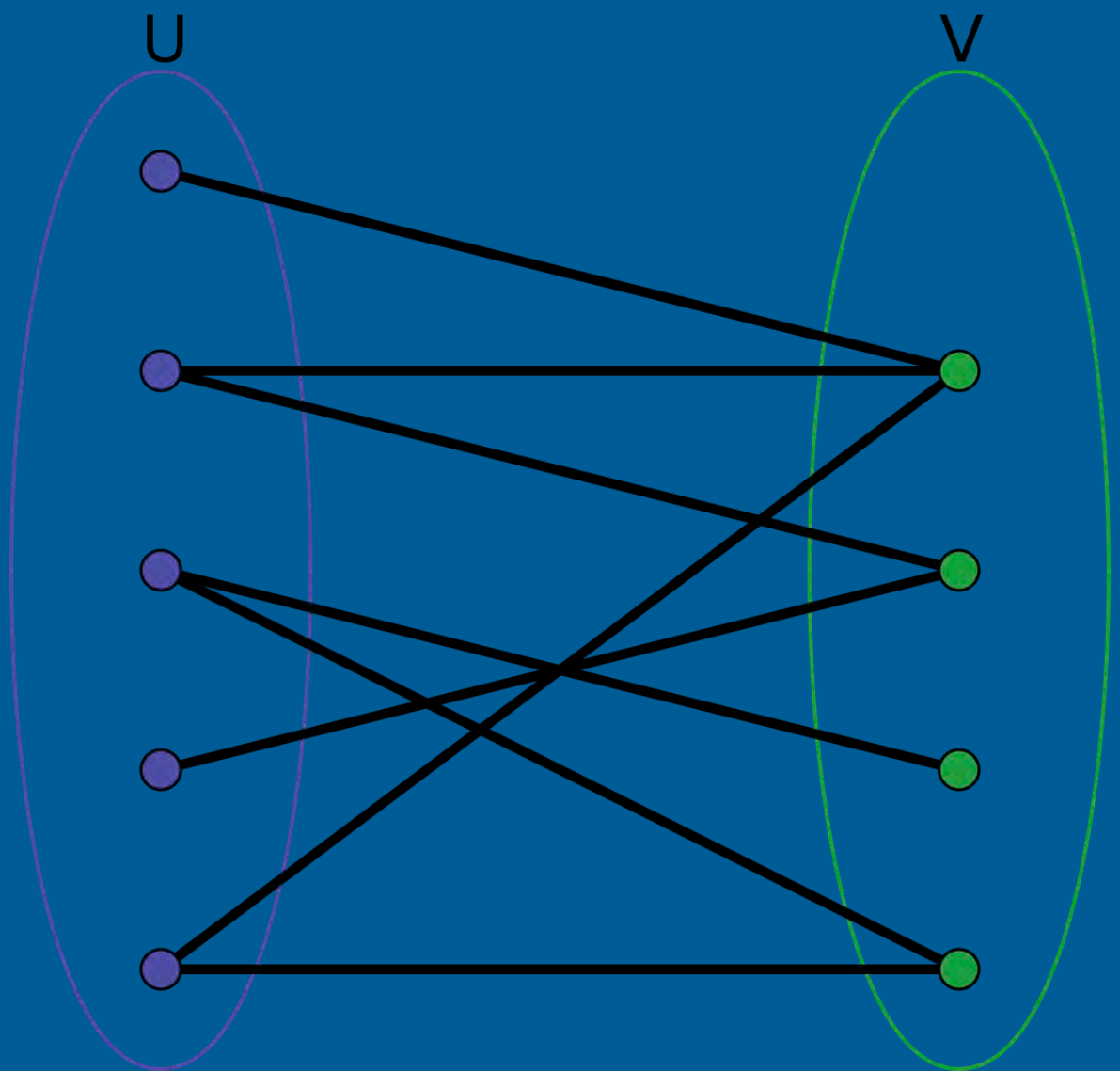


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.

Two Mode / Multipartite Networks

- Relations among two (or more!) different sets of nodes
  - Can be two different sets of people
- Also used between nodes of different conceptual type
  - People vs. interests / events
  - Important for the analysis of weak ties – we will get into this shortly



Making a Graph: Two Basic Ways

Edges Only

Edges + Nodes

Edges Only

Pros

- Quick and dirty
- Can get network stats fast
- Less data overhead to deal with
- Can specify connections

Cons

- No information about nodes
- Hard to filter / query
- Stuck with one data relationship

Edges + Nodes

Pros

- Can include far more information
- Can specify connections
- Data can be used elsewhere

Cons

- More information overhead
- Separation of network data
- Longer process to make a network

First Networks

Start with just edges in

Python

Python

Python Code | Result

```
 1 import networkx as nx
 2 import csv
 3 import pandas as pd
 4 from community import community_louvain
 5 import matplotlib.cm as cm
 6 import matplotlib.pyplot as plt
 7
 8 # First, get the edge data
 9 url = 'https://raw.githubusercontent.com/mathbeveridge/gameofthrones/master/data/got-s1-edges.csv'
10 dfedges = pd.read_csv(url)
11 dfedges
```

Python Graphs

Python Code | Result

```
1  import networkx as nx
2  import csv
3  import pandas as pd
4  from community import community_louvain
5  import matplotlib.cm as cm
6  import matplotlib.pyplot as plt
7
8  # First, get the edge data
9  url = 'https://raw.githubusercontent.com/mathbeveridge/gameofthrones/master/data/got-s1-edges.csv'
10 dfedges = pd.read_csv(url)
11 # Dataset is now stored in a Pandas Dataframe
12
13 # Now we create the graph from the edge list. We need to specify the column names as they are in mix
14 G = nx.from_pandas_edgelist(dfedges, source="Source", target = "Target", edge_attr=True)
15 # Draw the graph!
16 pos = nx.spring_layout(G, k=1, iterations=20)
17 nx.draw(G, pos, with_labels=True)
```

Python: Nodes

Python Code | Result

```
1  import networkx as nx
2  import csv
3  import pandas as pd
4  from community import community_louvain
5  import matplotlib.cm as cm
6  import matplotlib.pyplot as plt
7
8  # First, get the edge data
9  url = 'https://raw.githubusercontent.com/mathbeveridge/gameofthrones/master/data/got-s1-edges.csv'
10 dfedges = pd.read_csv(url)
11
12 # Now get the node data
13 urlNode = 'https://raw.githubusercontent.com/mathbeveridge/gameofthrones/master/data/got-s1-nodes.cs
14 dfnodes = pd.read_csv(urlNode)
15
16 # Dataset is now stored in a Pandas Dataframe
17
18 # Now we create the graph from the edge list. We need to specify the column names as they are in mix
19 G = nx.from_pandas_edgelist(dfedges, source="Source", target = "Target", edge_attr=True)
20 data = dfnodes.set_index('Id').to_dict('index').items()
21 G.add_nodes_from(data)
22
23 print(G.nodes(data=True))
```

Why Not Just Combine the Lists Later?

- You *can*, but…
  - You can use node attributes for styling
  - You can use node attributes for filtering
  - Many other uses

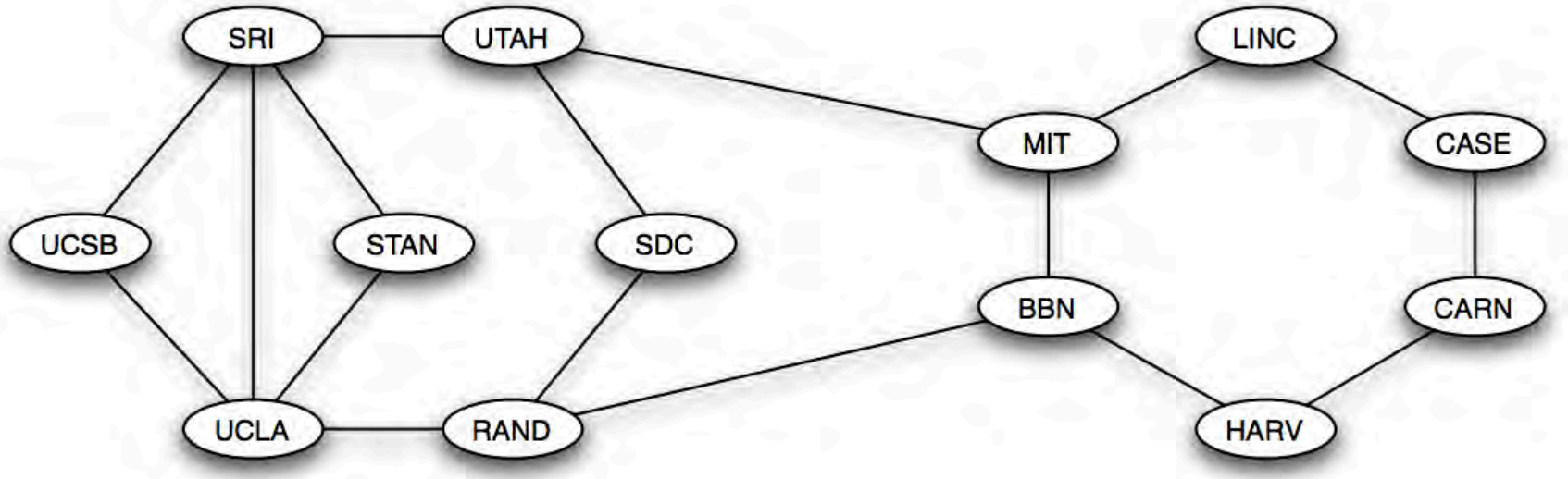
Exploring our Graph: Paths

- Movement in a graph via edges
- Sequence of nodes connected via edges
- *Simple path*: A path that does not repeat nodes

Connectivity

- Path between every pair of nodes
- Goal of most designed networks
- NOT a necessary feature of graphs though!
- There are social networks with disconnected features

Example – Anyone know what this is?



Graph Distance

- Not geographic (mostly!)
- Distance = length of the shortest path between two nodes
  - Number of edges
  - Sometimes we can simply look at this
  - Other times…we need computers!

Python: Graph Distance

Python Code | Result

```
path = dict(nx.all_pairs_shortest_path(G))
path
```

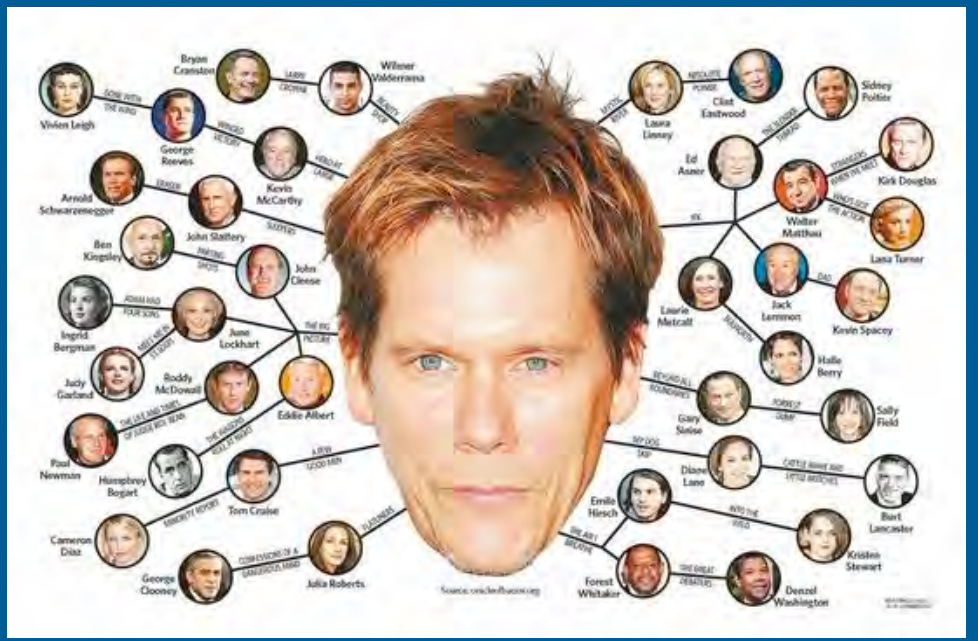
Python: Paths

Python Code | Result

```
1  import networkx as nx
2  import csv
3  import pandas as pd
4  from community import community_louvain
5  import matplotlib.cm as cm
6  import matplotlib.pyplot as plt
7
8  # First, get the edge data
9  url = 'https://raw.githubusercontent.com/mathbeveridge/gameofthrones/master/data/got-s1-edges.csv'
10 dfedges = pd.read_csv(url)
11 # Dataset is now stored in a Pandas Dataframe
12
13 # Now we create the graph from the edge list. We need to specify the column names as they are in mix
14 G = nx.from_pandas_edgelist(dfedges, source="Source", target = "Target", edge_attr=True)
15 # Draw the graph!
16 pos = nx.spring_layout(G, k=4, iterations=20)
17 nx.draw(G, pos, with_labels=True, node_size=5, font_size=5, width=.2)
18 # draw path in red
19 path = nx.shortest_path(G,source='HUGH_OF_THE_VALE',target='MIRRI_MAZ_DUUR')
20 path_edges = list(zip(path,path[1:]))
21 nx.draw_networkx_nodes(G,pos,nodelist=path,node_color='r')
22 nx.draw_networkx_edges(G,pos,edgelist=path_edges,edge_color='r',width=10)
23 plt.axis('equal')
24 plt.show()
```

Small-world phenomenon

- Shorter path then you would think to get from one node to another
- Origin of the term “six degrees of separation”
- Practical terms: Who has a friend from another country?



Components

- Natural breaks for connected portions of a graph
- Connected component of a graph
  - Every node in the subset has a path to each other

Giant Component

Informal definition:
Connected component that contains a significant fraction of all the nodes

- Return to Game of Thrones: People who want Joffrey dead. Not **everybody**, but close!
- Most networks only have one giant component

Python: Giant Component

Python Code | Result

```
1  # Draw the graph!
2  pos = nx.spring_layout(G, k=1, iterations=20)
3  Gcc = sorted(nx.connected_components(G), key=len, reverse=True)
4  G0 = G.subgraph(Gcc[0])
5  nx.draw(G0, pos, with_labels=True)
```

Merger of Giant Components

- Only one connection merges giant components into one
  - In history: Sudden, often catastrophic change
  - Think of 1492 C.E.
    - Disease
    - Political change
    - Previous contacts were not sustained
- Issue of time

Beyond Visualization

- You know about *nodes*, *edges*, *paths*, *distance*, *components*, *directed* and *undirected* graphs
- Time to move on from description of graphs to some operations
- These are some of the most common measurements

Network Density

- Number of actual edges divided by total number of possible edges
  - Actual vs. potential connections
- One of the most basic measurements in SNA
- Especially interesting in epidemiology
  - Or any analysis of how something spreads in a network

Python: Network Density

Python Code | Result

```
nx.density(G)
```

Degree Measurements

- Sum of all other nodes with a direct *path* to a node
- Signifies activity or popularity
- Very good for looking at nodes in a local context
- In, Out
  - Weighted and unweighted

Centrality

- *VERY* important concept!
- What many people want to see in networks: the most prominent nodes
- These are often the “key” players in a network
- Idea of social power
  - Assertion: Power is inherently relational

Measurements

Degree

Betweenness

Closeness

Eigenvector

Python: Degree Measurements

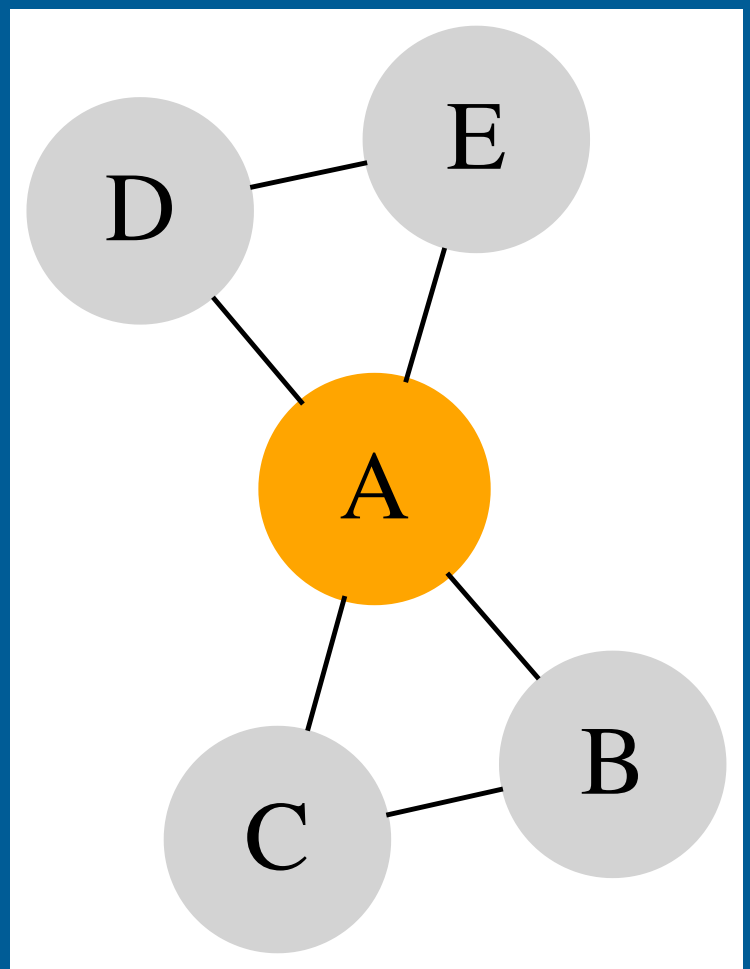
Python Code | Result

```
G.degree
```

Size by Degree

- Now we will visualize this on our graph
- Same techniques can be used for all other measurements / attributes

Degree



Python: Degree graph

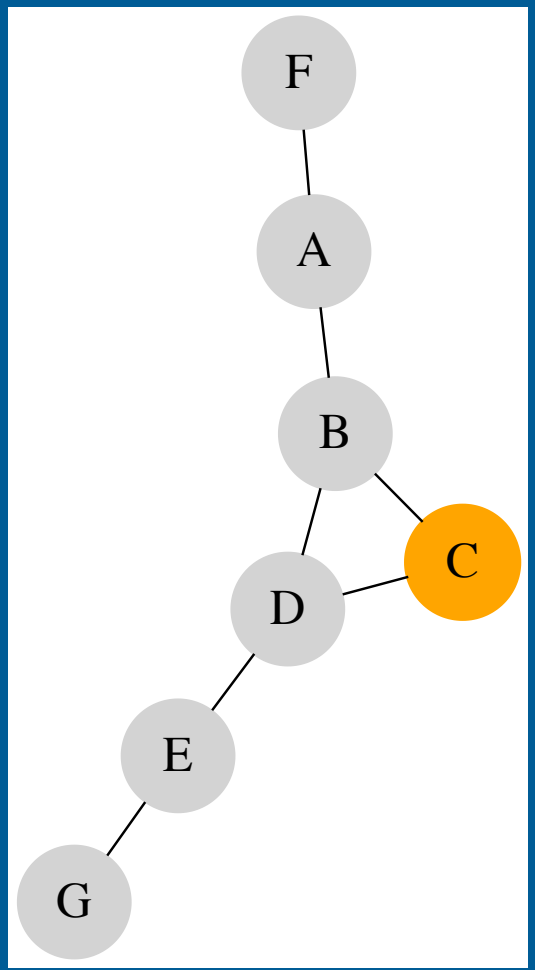
Python Code | Result

```
1  d = dict(G.degree)
2
3  nx.draw(G, nodelist=d.keys(), node_size=[v * 10 for v in d.values()])
4  plt.show()
```

Betweeness Centrality

- Measures how often a node appears on shortest paths between nodes in the network
- Often better to change visualization to identify them
- A node with a high measure here could be important…or at the periphery of multiple networks

Betweeness Centrality



Python: Betweeness Centrality

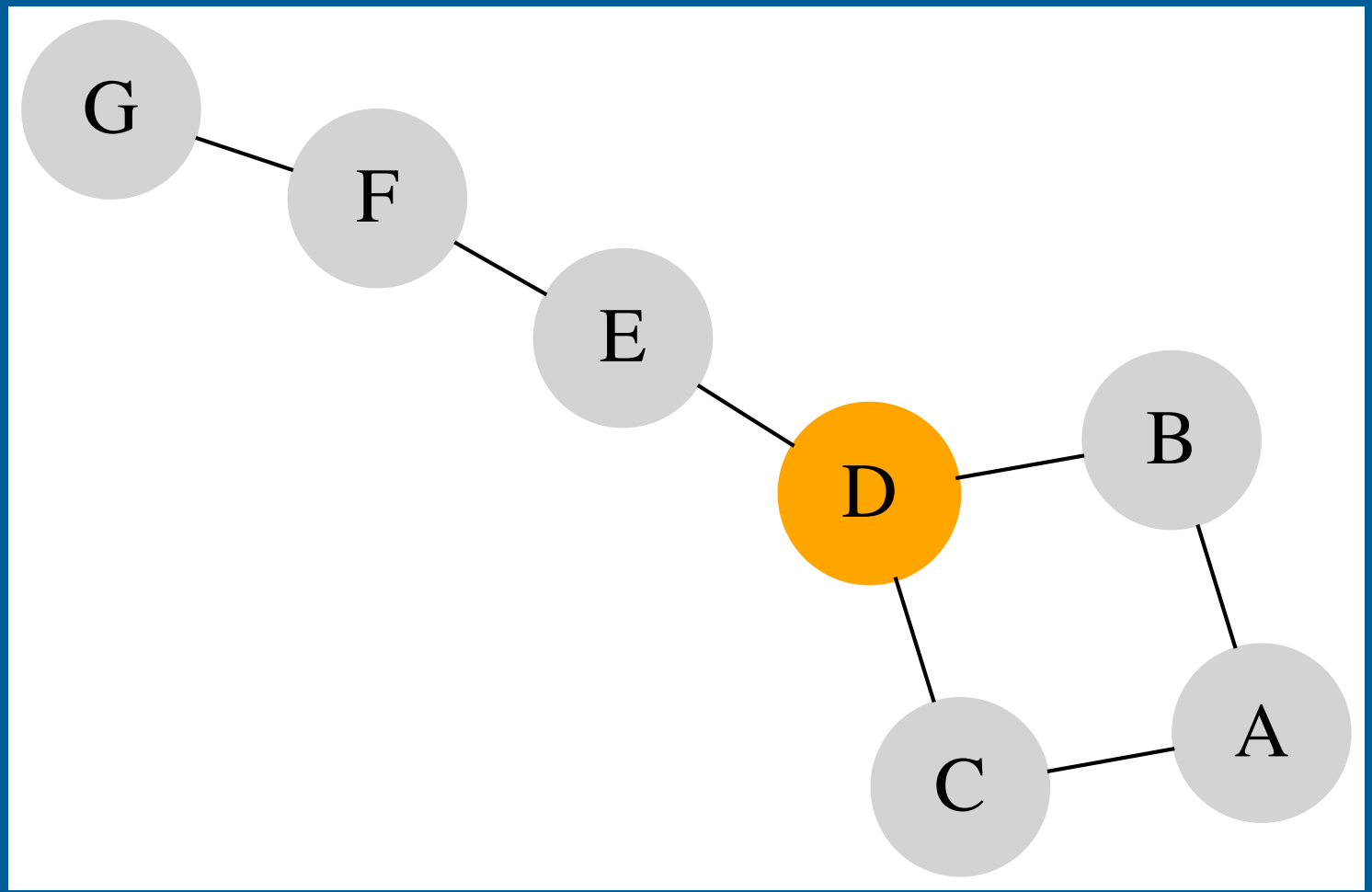
Python Code | Result

```
1  nx.betweenness_centrality(G)
```

Closeness Centrality

- A sum of the shortest path between all nodes from a node
- Useful to find out who spreads information quickly; might not be useful in a highly connected network

Closeness Centrality



Python: Closeness Centrality

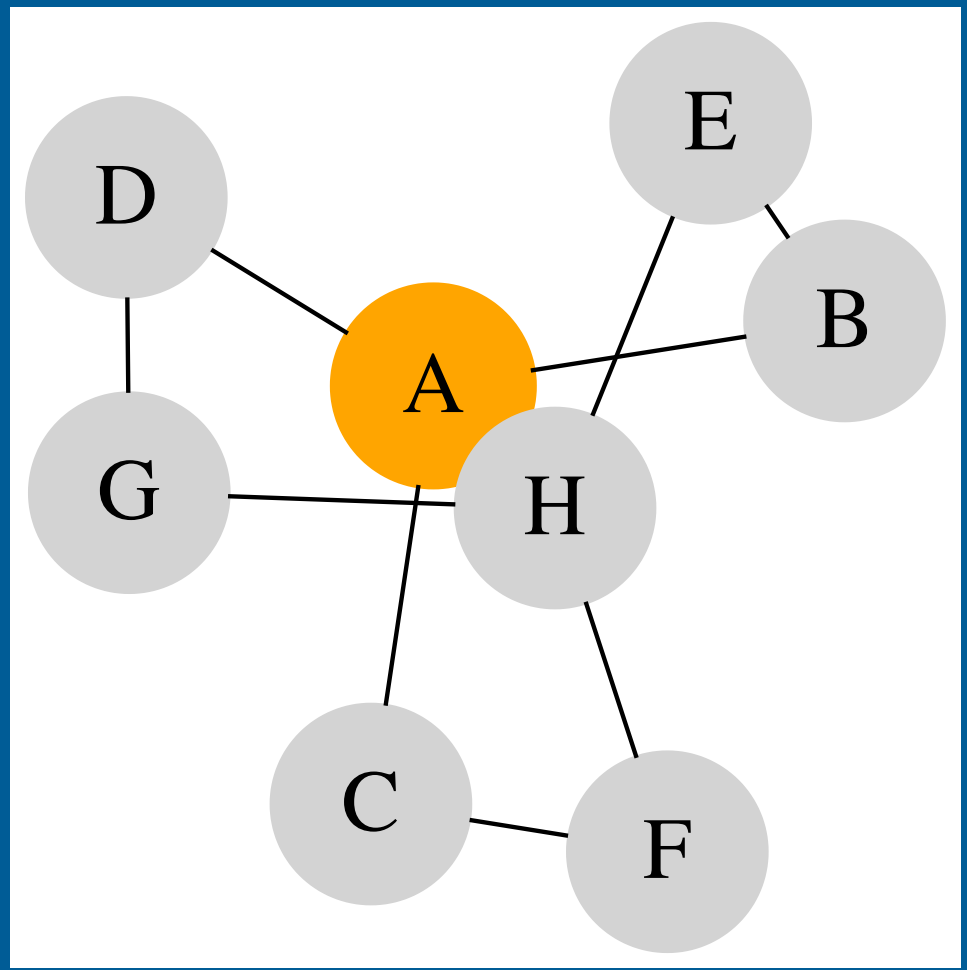
Python Code | Result

```
nx.closeness_centrality(G)
```

Eigenvector Centrality

- Basic idea: A node is important if it is linked to by other important nodes
  - I have almost no friends - but my friends are
    - The President
    - The Pope
    - Santa Claus
    - *Insert influencer / Reality TV star here*
- Global vs. local importance
- Mathematical modeling of this stretches back to the 1940s
  - Software does this for us!

Eigenvector Centrality



Python: Eigenvector Centrality

Python Code | Result

```
nx.eigenvector_centrality(G)
```

Page Rank

- Similar to Eigenvector, but takes into account the direction and weight of connections
- Famously used by Google

Python: Page Rank

Python Code | Result

```
nx.pagerank(G)
```

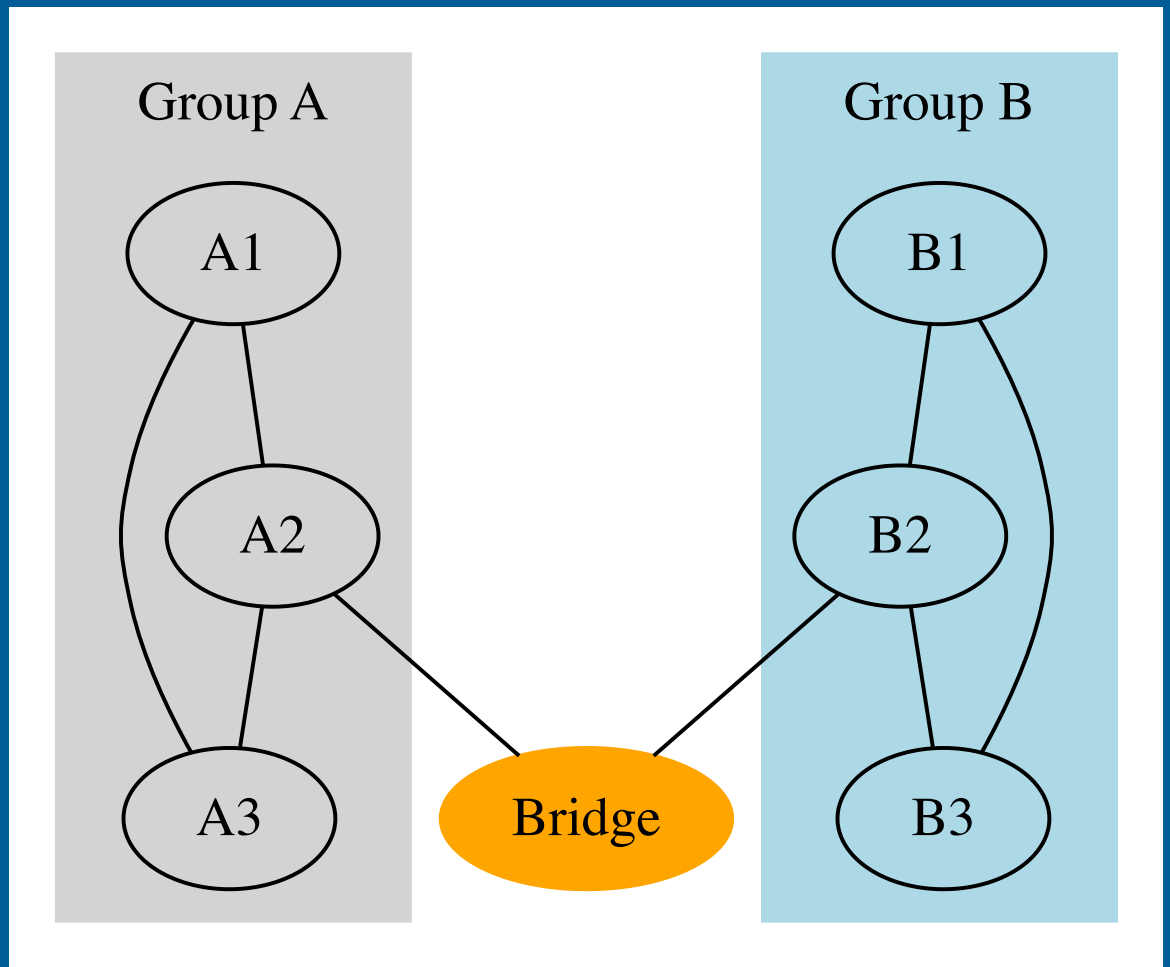
Centrality Summary

- **Degree**: Most direct connections.
- **Betweenness**: Bridges many paths.
- **Closeness**: Nearest on average to others.
- **Eigenvector**: Connected to important others.

Bridges

- Edge joining two nodes is a bridge if deleting that edge would cause the nodes to be in two different components of the graph
  - Very rare in social networks – think of the small world principle
- Local bridge if the endpoints have no friends in common
  - There can still be a path between the nodes if the bridge is deleted
- Best statistical approximation: Betweenness Centrality

Bridges



Python: Bridges

Python Code | Result

```
nx.has_bridges(G)
```

Python: Bridges

Python Code | Result

```
list(nx.bridges(G))
```

Python: Highlight Bridges

Python Code | Result

```python
# Identify the bridges
bridges = list(nx.bridges(G))

# Set positions for nodes using a layout algorithm
pos = nx.spring_layout(G, k=.01, iterations=20)

# Draw entire network
nx.draw(G, pos, with_labels=False, node_size=5, width=.2)

# Draw bridges in red
for bridge in bridges:
    nx.draw_networkx_edges(G, pos, edgelist=[bridge], edge_color='red', width=2)
    nx.draw_networkx_nodes(G, pos, nodelist=bridge, node_color='red', node_size=700)

# Create a dictionary of labels for nodes involved in bridges
bridge_labels = {node: node for bridge in bridges for node in bridge}

# Draw labels for nodes involved in bridges
nx.draw_networkx_labels(G, pos, labels=bridge_labels, font_size=8, font_weight='bold')

plt.show()
```

Connection Strength

- No real formal, strict definition
- Stronger links represent closer friendship / greater degrees of interaction
- Largely up to you!

Strong Ties and Weak Ties

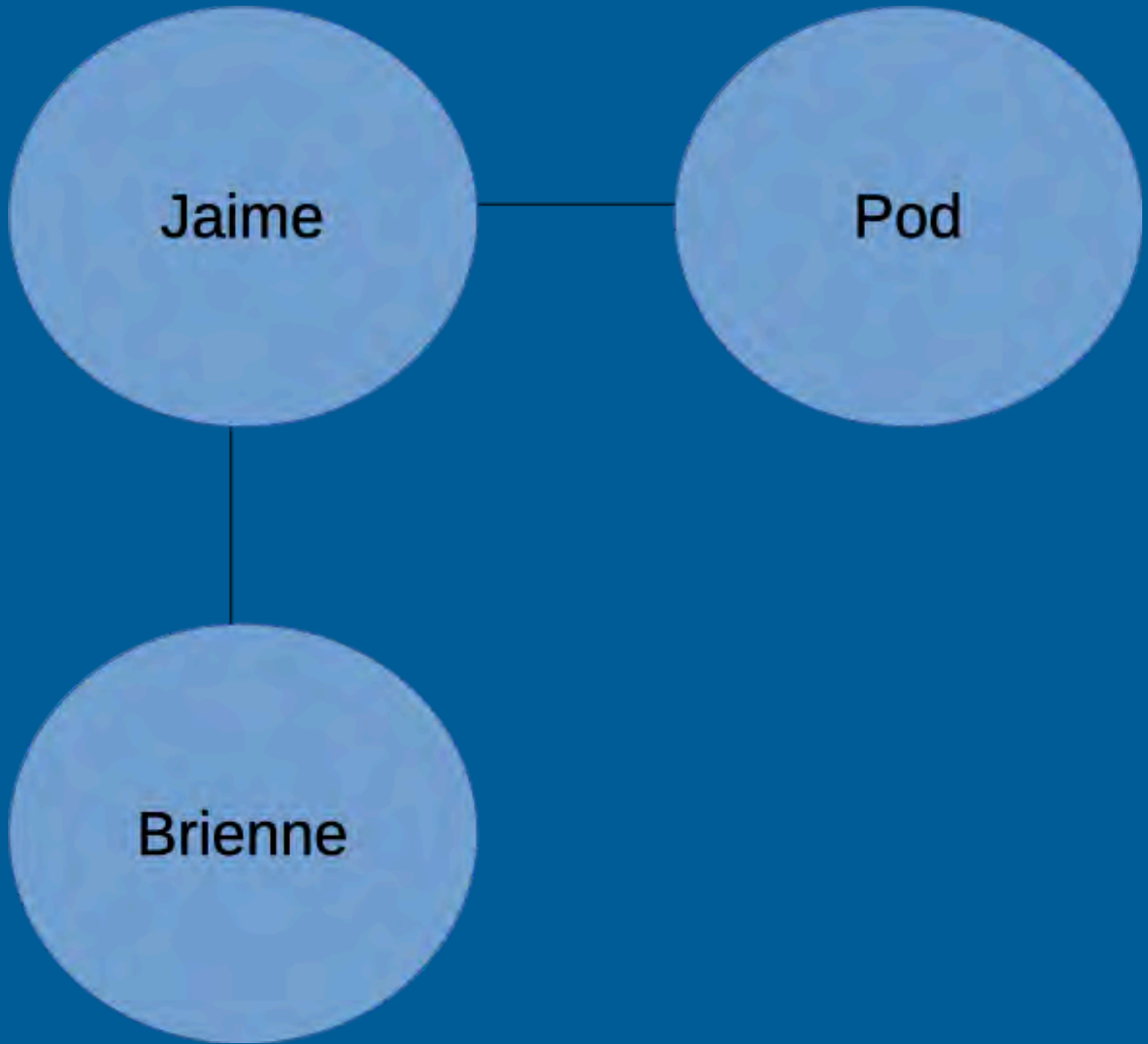
- Strong ties = close friendship, etc
- Weak ties = acquaintances, etc

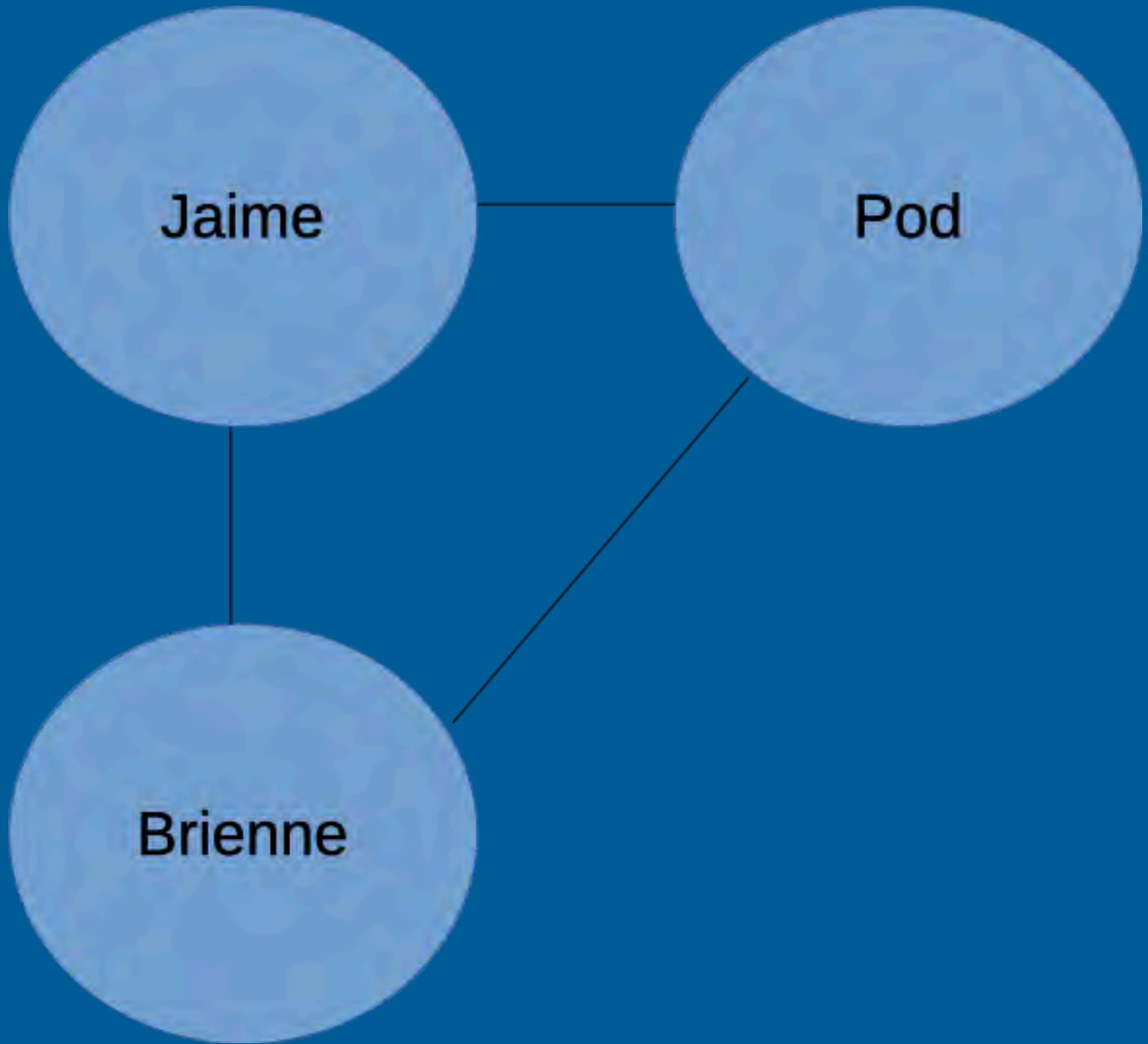
Network Evolution: Triadic Closure

- Concept: If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future
- Example: If Jaime and Pod are friends, and Jaime and Brienne of Tarth are friends, then it is likely that Brienne and Pod will become friends

Triadic Closure

- This forms a triangle in the graph
- Generally form in social networks given enough time
- Think facebook friends feature
- Basic functionality of a social network





Python: Find All Potential Triads

Python Code | Result

```
plt.savefig('graph1.png')
```

Position at Structural Holes

- Structural hole
  - Empty space between sets of nodes that do not otherwise interact closely
- Position is advantageous
  - Access to information that other components might not have
  - Energy in reaching out to different groups rather than the same one

Other Issues

- Synthesis from multiple ideas
- Social gatekeeping
  - A source of power as the position regulates communications
  - Some people might try to keep triangles from forming!
- Interests of a node and an organization may not be aligned
  - Connection may not be in the node's best interests
  - Unknown amount of time before triadic closure happens

Community Detection

- The number of connections between nodes is more dense in a certain grouping than outside it
- GOOD IDEA / common practice to make this the color of your nodes!

Python: Community Detection

Python Code | Result

```
partition = community_louvain.best_partition(G)
print(partition)
```

Python: Community Detection

Python Code | Result

```
partition = community_louvain.best_partition(G)
pos = nx.spring_layout(G)
# color the nodes according to their partition
cmap = cm.get_cmap('viridis', max(partition.values()) + 1)
nx.draw_networkx_nodes(G, pos, partition.keys(), node_size=40,
                       cmap=cmap, node_color=list(partition.values()))
nx.draw_networkx_edges(G, pos, alpha=0.5)
plt.show()
```

Thank You!

Any Questions?