Carnegie Mellon University

Uncertainty, m ss ng information, and network analysis

Matthew Lincoln, PhD
Research Software Engineer
@matthewdlincoln

Uncertainty is like the weather

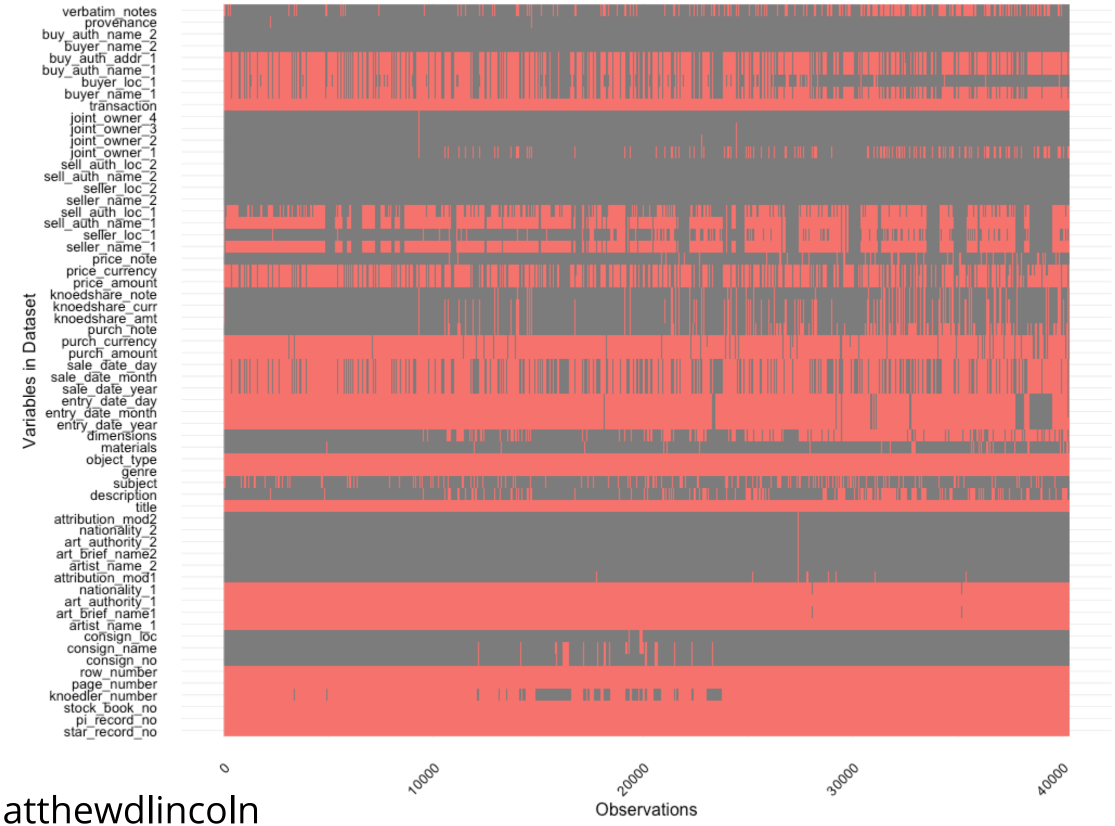
Everyone talks about it but nobody *does* anything about it.

Uncertainty is like the weather

~~Everyone talks about it but nobody *does* anything about it.~~

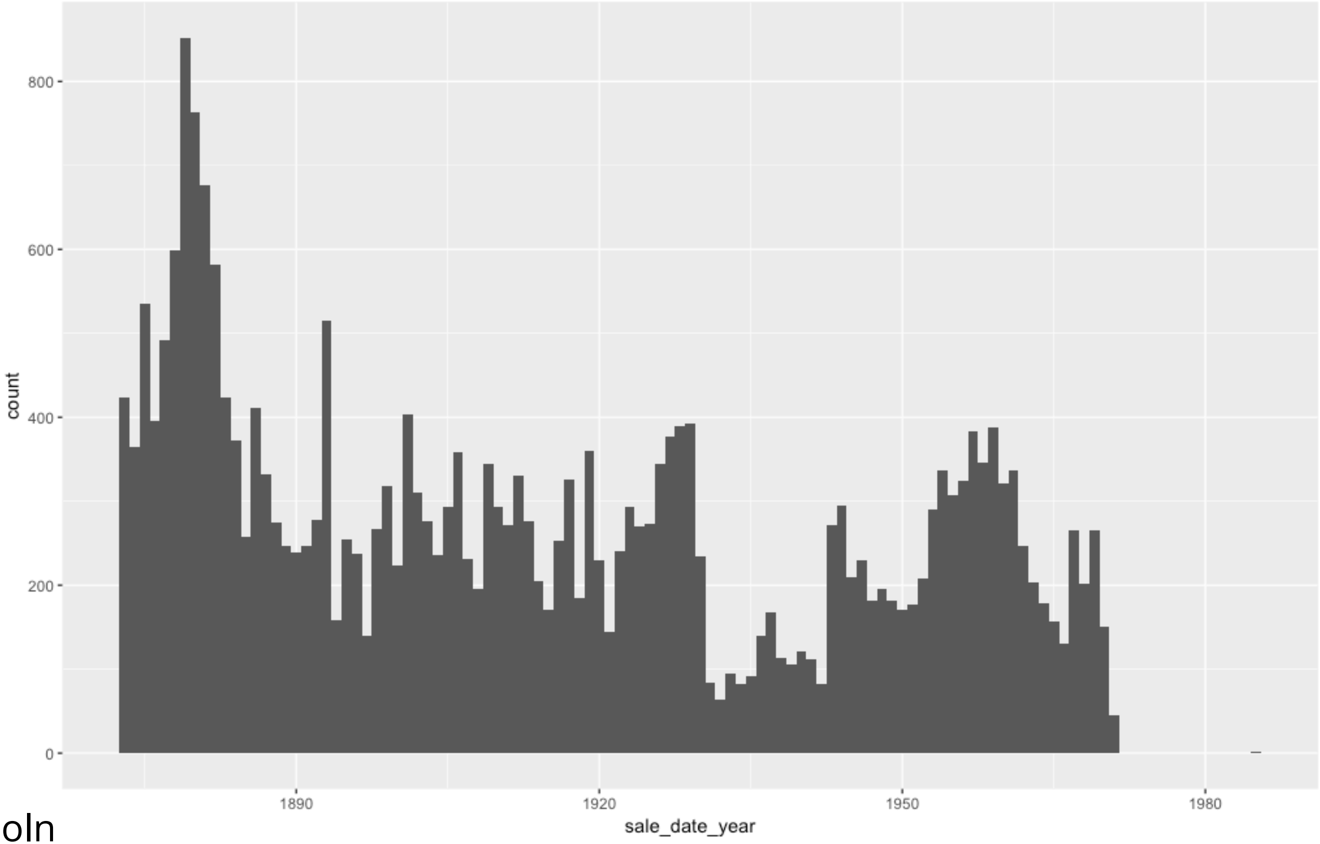
1. Ignore it
2. Constrain our conclusions
    - Document our uncertainty
    - But this is hard to do in networks!
3. Simulate our uncertainty

Ignore it



@matthewdlincoln

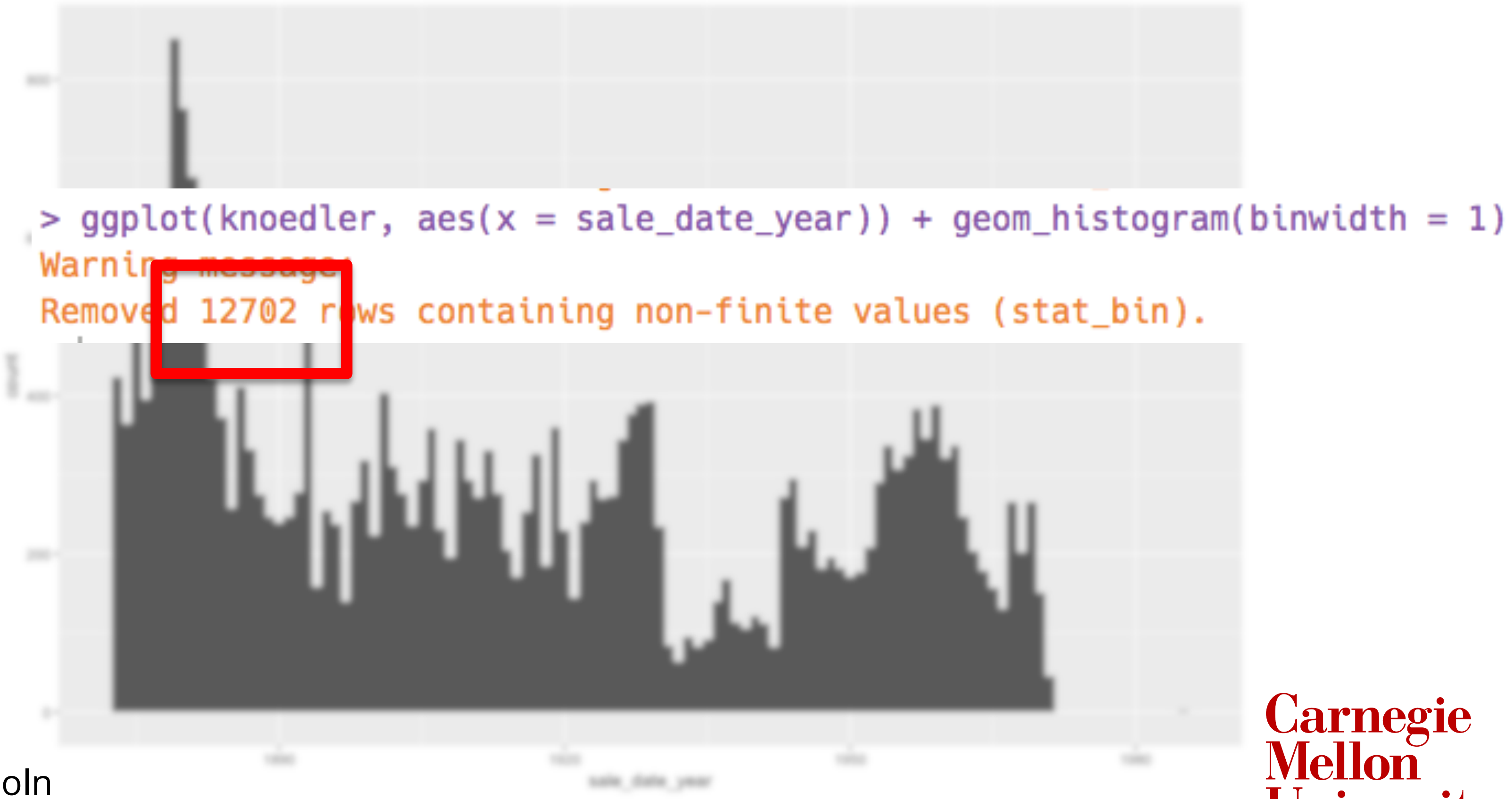
Ignore it



@matthewdlincoln

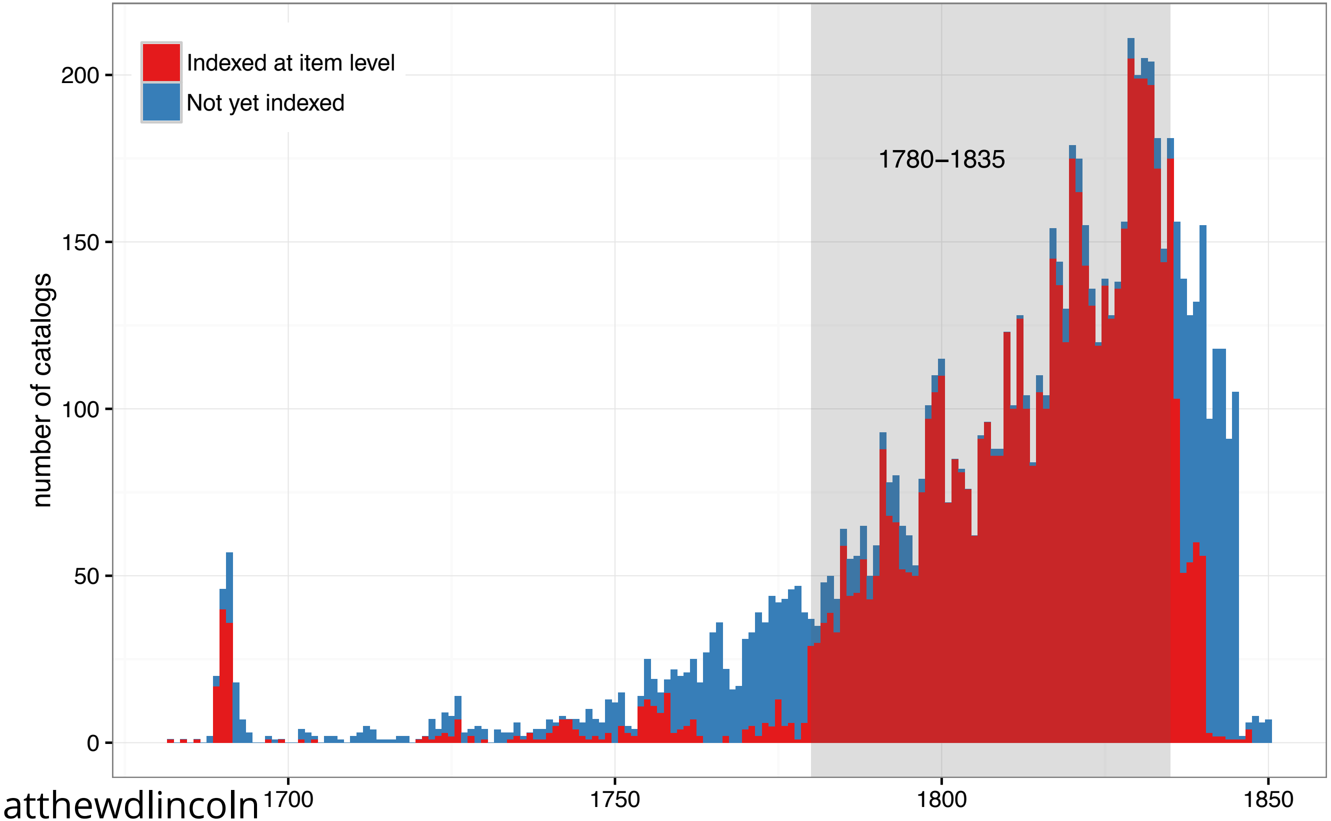
Carnegie Mellon University

Ignore it



```
> ggplot(knoedler, aes(x = sale_date_year)) + geom_histogram(binwidth = 1)
Warning message:
Removed 12702 rows containing non-finite values (stat_bin).
```

Constrain our results



Matthew Lincoln and Abram Fox, “The Temporal Dimensions of the London Art Auction, 1780-1835,” *British Art Studies 4* (Fall 2016)

@matthewdlincoln

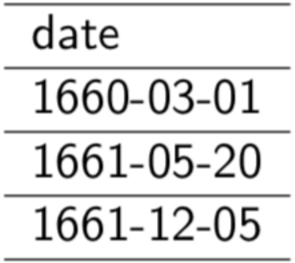
Carnegie Mellon University

Document our uncertainty

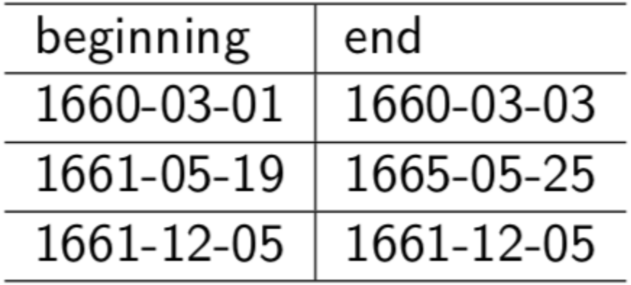
Dates

1. “point” events? Or durations?
2. How precise are your sources?
3. How varied is that precision?

Document our uncertainty

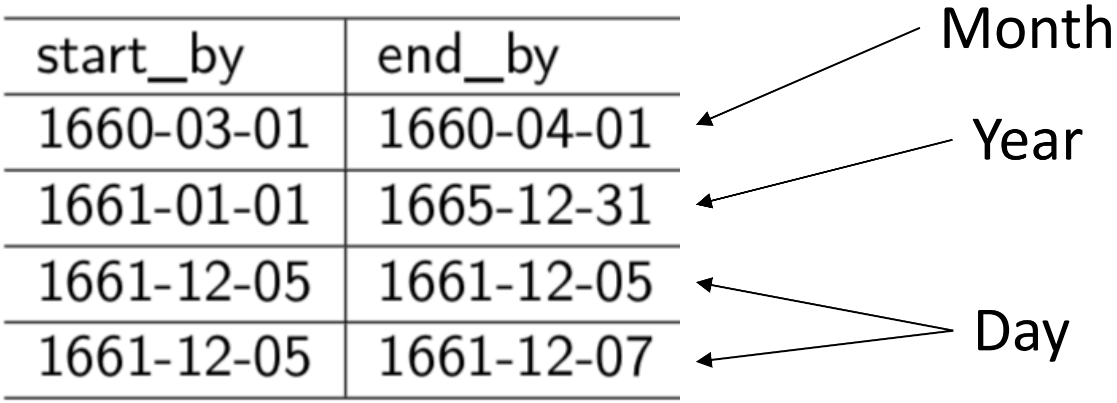


date
1660-03-01
1661-05-20
1661-12-05



beginning	end
1660-03-01	1660-03-03
1661-05-19	1665-05-25
1661-12-05	1661-12-05

Document our uncertainty



start_by	end_by
1660-03-01	1660-04-01
1661-01-01	1665-12-31
1661-12-05	1661-12-05
1661-12-05	1661-12-07

Month

Year

Day

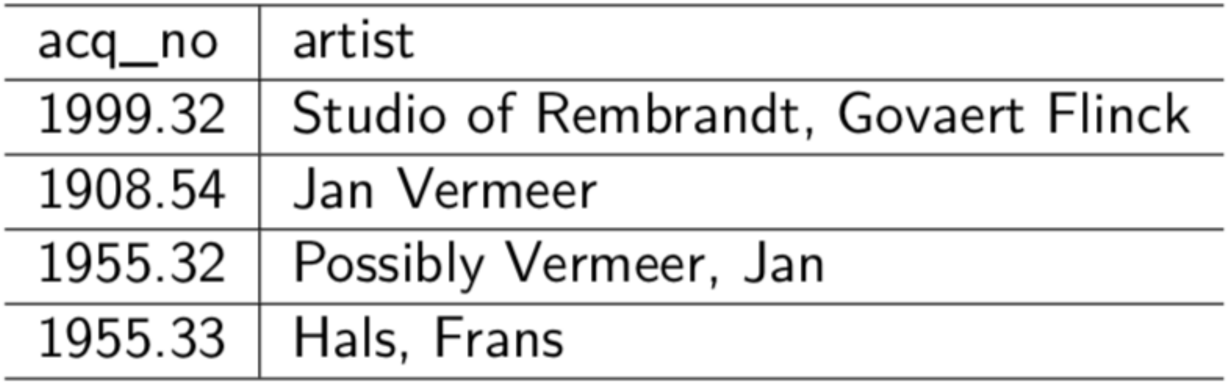
Document our uncertainty

- If you’re dealing with times, not just dates. . . then watch out for time zones. Python and R both have specialized libraries for these.
- When hand entering dates, make sure to validate the dates! You will inevitably enter YYYY-02-31, which doesn’t exist.

Document our uncertainty

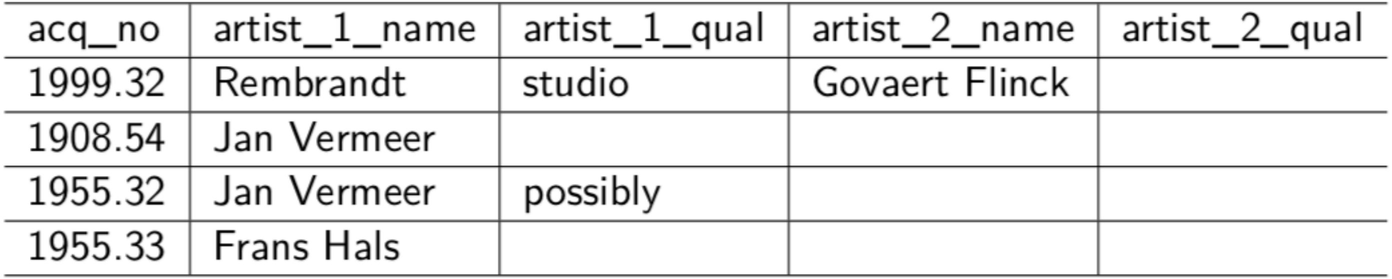
- What to do with uncertain data?
  - I need point data: take the midpoint of date ranges
  - Temporal network analysis: use start & end dates to establish when edges or nodes enter / leave the network
    - https://programminghistorian.org/en/lessons/temporal-network-analysis-with-r
  - Randomly sample within date range (more on that later)

Document our uncertainty



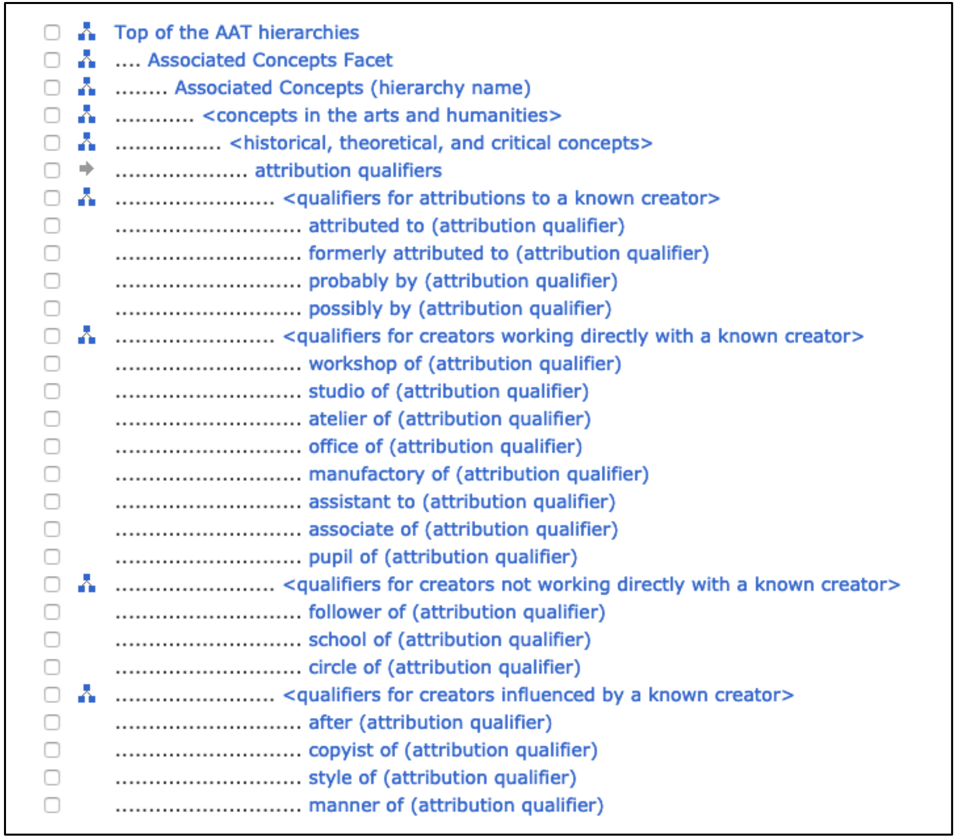
acq_no	artist
1999.32	Studio of Rembrandt, Govaert Flinck
1908.54	Jan Vermeer
1955.32	Possibly Vermeer, Jan
1955.33	Hals, Frans

Document our uncertainty



acq_no	artist_1_name	artist_1_qual	artist_2_name	artist_2_qual
1999.32	Rembrandt	studio	Govaert Flinck	
1908.54	Jan Vermeer			
1955.32	Jan Vermeer	possibly		
1955.33	Frans Hals			

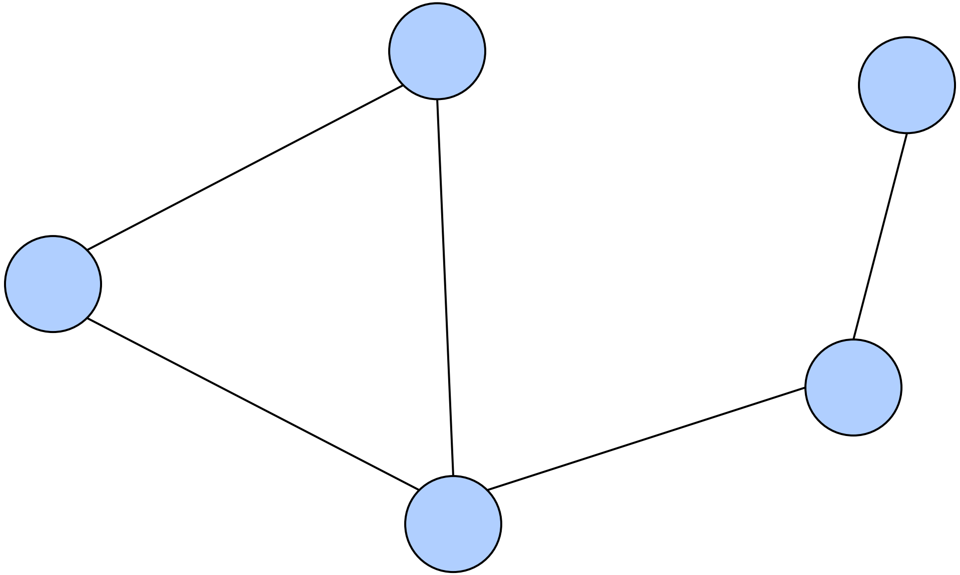
Document our uncertainty



□ Top of the AAT hierarchies
□ .... Associated Concepts Facet
□ .... Associated Concepts (hierarchy name)
□ .... <concepts in the arts and humanities>
□ .... <historical, theoretical, and critical concepts>
□ .... attribution qualifiers
□ .... <qualifiers for attributions to a known creator>
□ .... attributed to (attribution qualifier)
□ .... formerly attributed to (attribution qualifier)
□ .... probably by (attribution qualifier)
□ .... possibly by (attribution qualifier)
□ .... <qualifiers for creators working directly with a known creator>
□ .... workshop of (attribution qualifier)
□ .... studio of (attribution qualifier)
□ .... atelier of (attribution qualifier)
□ .... office of (attribution qualifier)
□ .... manufactory of (attribution qualifier)
□ .... assistant to (attribution qualifier)
□ .... associate of (attribution qualifier)
□ .... pupil of (attribution qualifier)
□ .... <qualifiers for creators not working directly with a known creator>
□ .... follower of (attribution qualifier)
□ .... school of (attribution qualifier)
□ .... circle of (attribution qualifier)
□ .... <qualifiers for creators influenced by a known creator>
□ .... after (attribution qualifier)
□ .... copyist of (attribution qualifier)
□ .... style of (attribution qualifier)
□ .... manner of (attribution qualifier)

AAT “attribution qualifiers” hierarchy

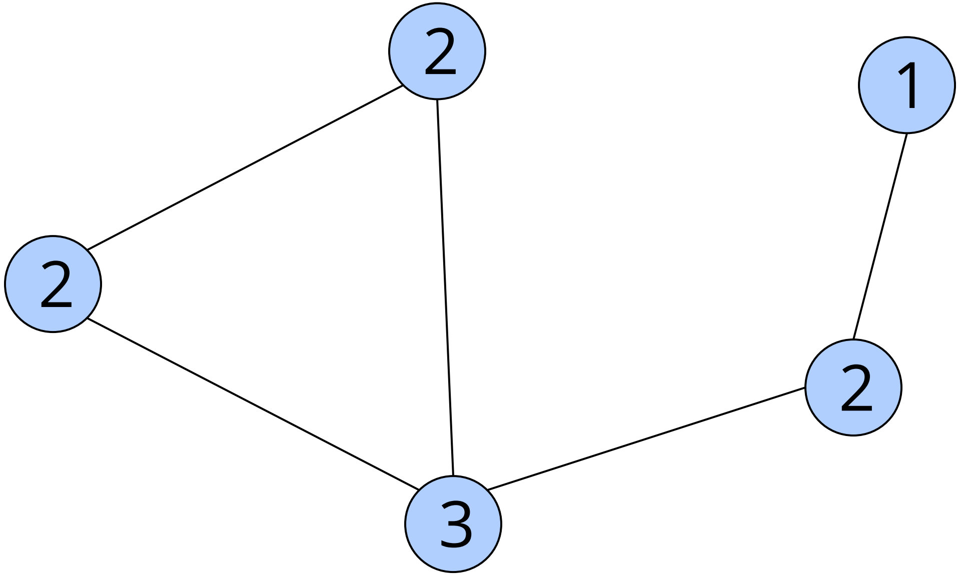
Missing data is “viral” in networks



@matthewdlincoln

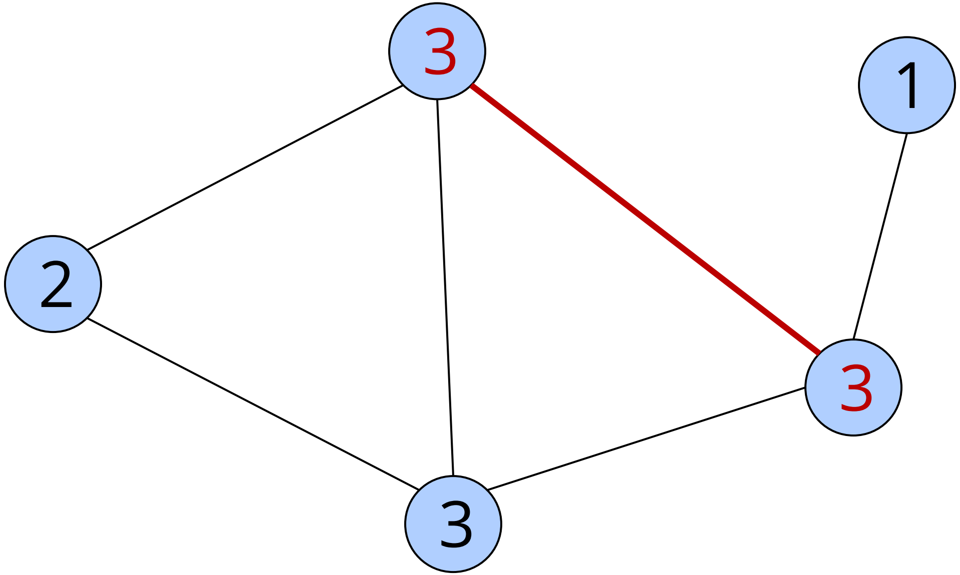
Missing data is "viral" in networks

Degree

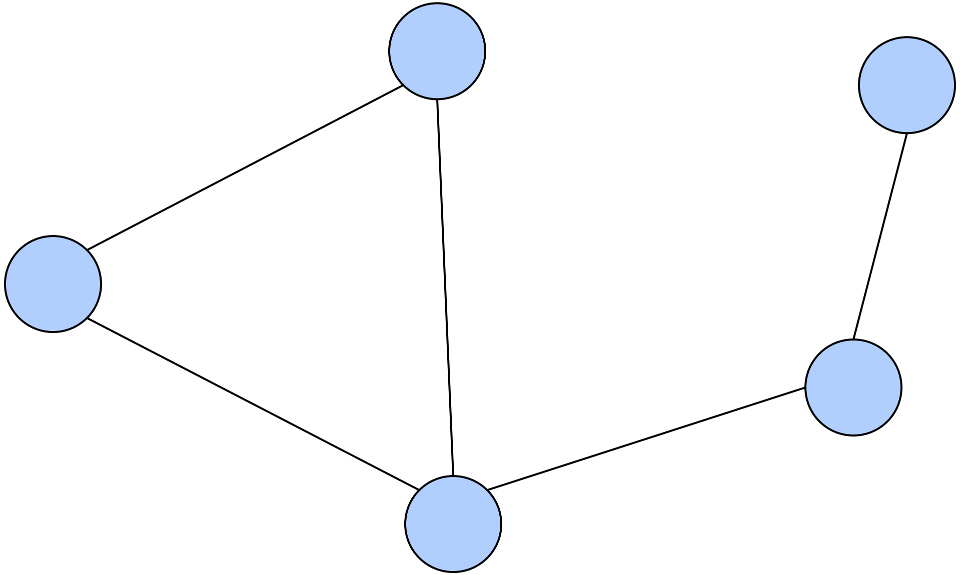


Missing data is “viral” in networks

Degree

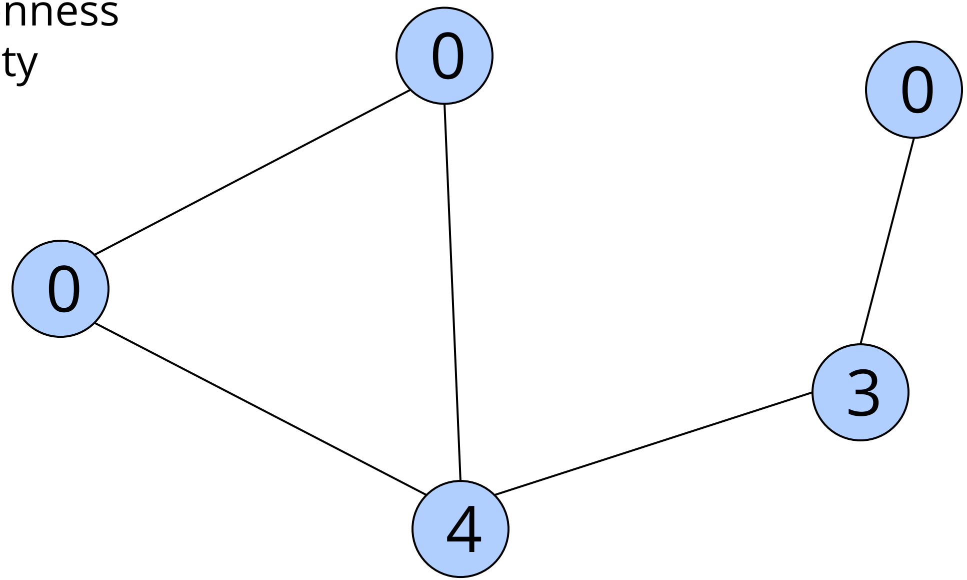


Missing data is "viral" in networks



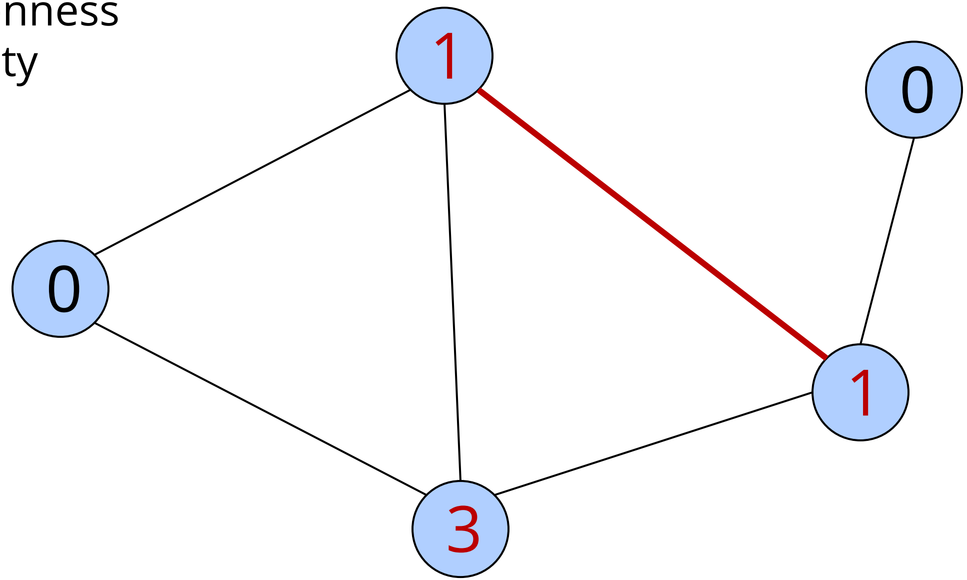
Missing data is “viral” in networks

Betweenness centrality



Missing data is “viral” in networks

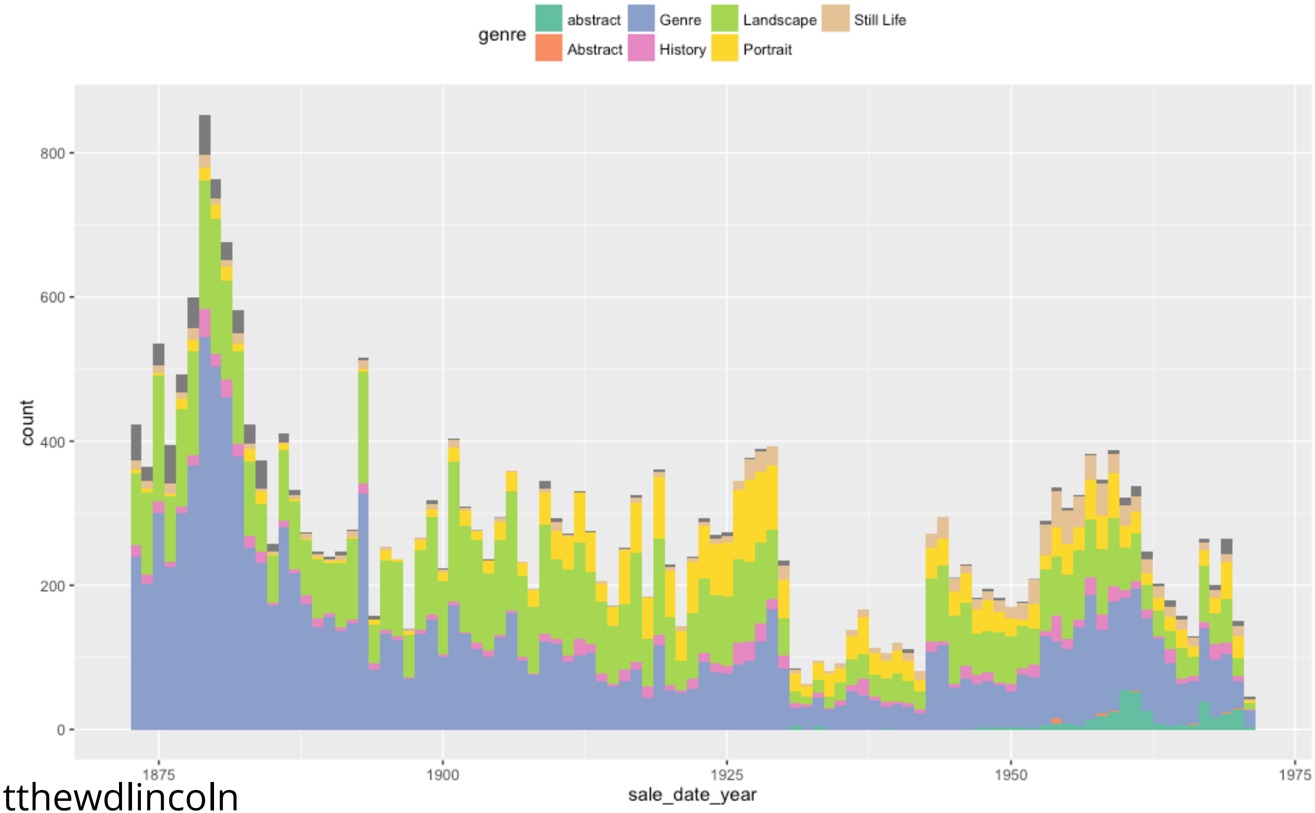
Betweenness centrality



Missing data is “viral” in networks

- Avoid metrics that rely on shortest path lengths (graph diameter, betweenness centrality)
- Metrics about the entire graph (density or connectivity) are slightly safer to use when you know you have missing data

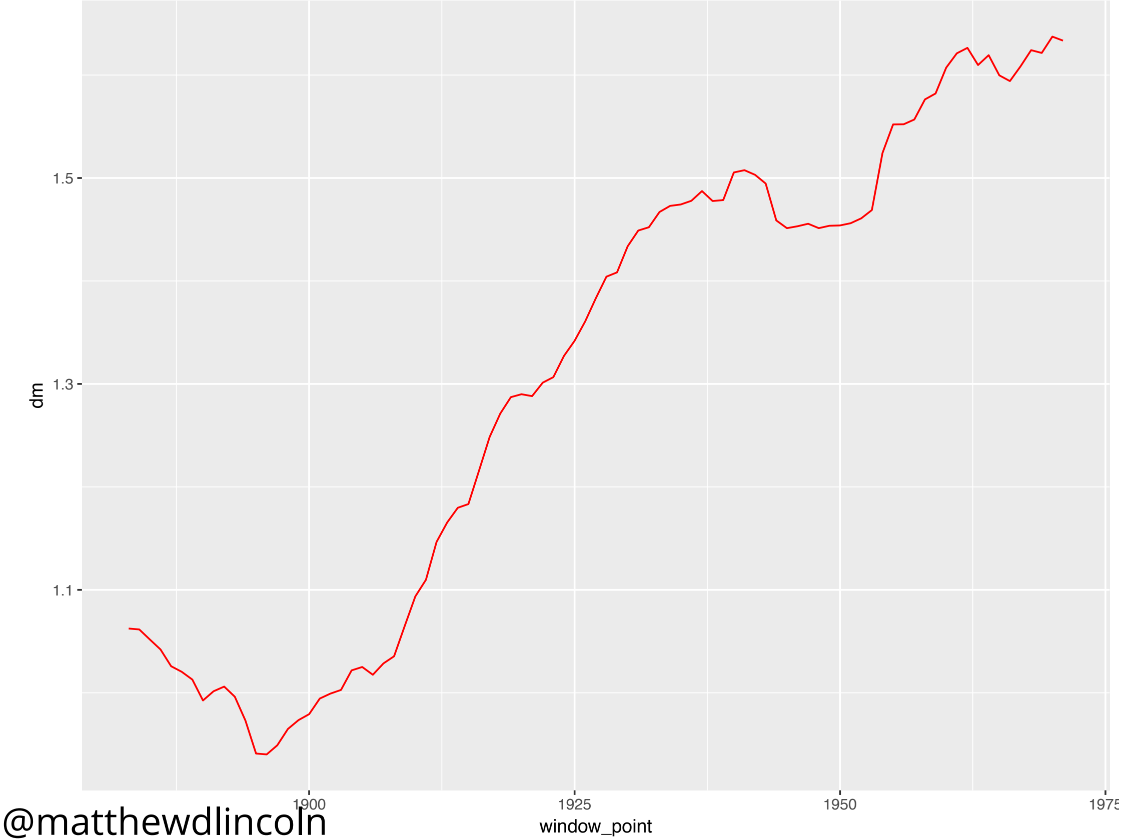
Simulate our uncertainty



Sales per year in the Knoedler Stockbooks, colored by genre

@matthewdlincoln

Simulate our uncertainty



Shannon diversity measure of genre (10-year rolling window)

Simulate our uncertainty

https://mdlincoln.shinyapps.io/missingness/

Think / Pair / Share

1. Get together with your group
2. Identify ONE variable in your data where you have uncertainty/missing data
3. Present this variable to the group: how are you managing missing values, **and how does it affect your research question?**