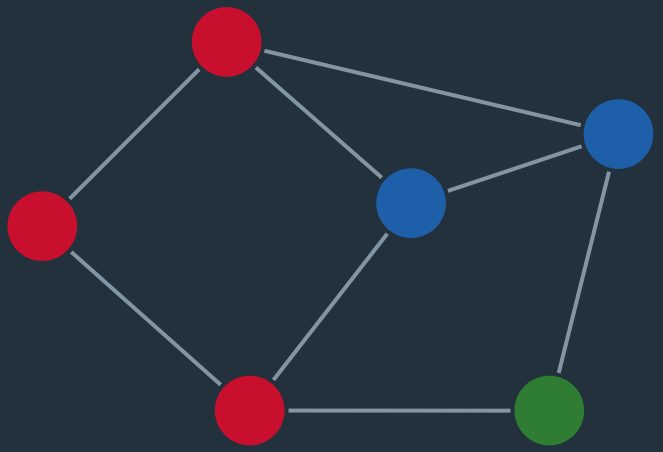
Networks as data

Course C · Workshop · Day 1

Chapters 1 to 5 of Network Analysis with Stata (Stata Press), about six hours

Instructors are free to use, adapt and share these slides.



Day 1: five blocks, two halves

9:00	**1 Welcome and set-up**		10 min
9:10	**2 Why networks?**	Chapter 1	40 min
9:50	**3 What is a network?**	Chapter 2	55 min
10:45	Break		15 min
11:00	**4 From a question to a network**	Chapter 3	50 min
11:50	Lunch		60 min
12:50	**5 Working with networks in Stata**	Chapter 4	85 min
14:15	Break		15 min
14:30	**6 Seeing networks**	Chapter 5	75 min
15:45	**7 Wrap-up and tomorrow**		15 min

Before we start: three lines that must work

```
. net install nwcommands, from(https://nwcommands.org/net/) replace
. nwinstall, everything
. nwversion
```

Installs the core and the four model packages; running the two install commands again updates everything.

```
. nwwebuse florentine, nwclear
. nwplot flomarriage
```

If this draws sixteen families, everything is in place.

Stata 17 or later

Any edition; Stata/BE runs every example of the book.

Internet in the room

nwwebuse loads the example data sets from nwcommands.org.

The do-files

nwcommands.org/book/dofiles: every printed command, chapter by chapter, to run alongside the slides.

One firm, two small examples, three ideas

The law firm

71 lawyers in three offices, three relations among them: advice, work together, friendship (Lazega). The running example of the book.

Four associates

Anna, Sophie, Gwen and Leh, typed in by hand: small enough to check every number.

Sixteen Florentine families

Marriage ties around 1430: the network used to learn what a picture does and does not show.

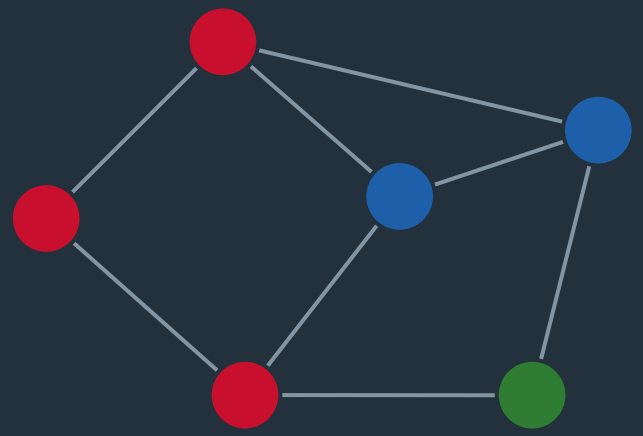
Three ideas for the day

- A relational question needs data on pairs, and every construction decision changes the network.
- A network is declared once in Stata and found by name afterwards.
- Node positions in a picture are not data.

1

Why networks?

40 min · Chapter 1 · no code in this block



A relational question needs data on pairs

	Attribute question	Relational question
Example	Do smokers tend to have lower income?	Does having a smoking friend predict starting to smoke, net of one's own background?
Data needed	One row per person, one column per characteristic	Who is friends with whom: data on pairs, in addition to the rows
Can ordinary regression answer it?	Yes	Not from rows alone: a limit of the data, not of the statistics

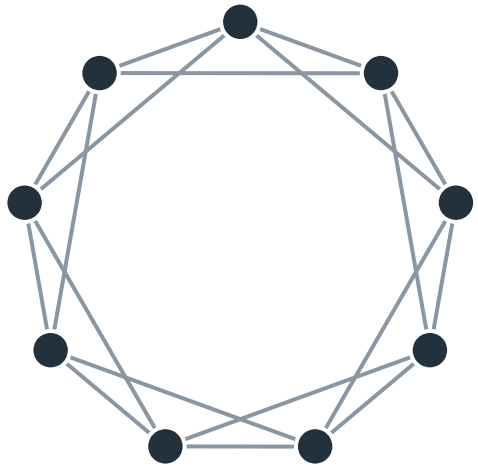
Attributes belong to one unit

Anna's age can be written down without mentioning anyone else. Friendship is a property of the pair (Anna, Sophie).

A count is not the network

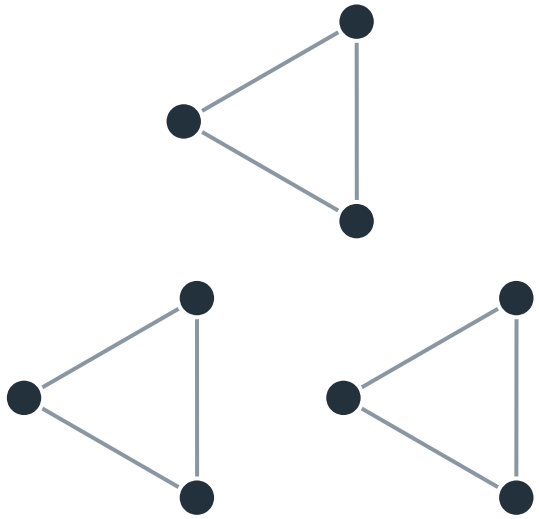
A firm's number of trading partners keeps one fact about the ties. It does not say who the partners are or whether they trade with each other.

Same nine people, three structures, three different futures



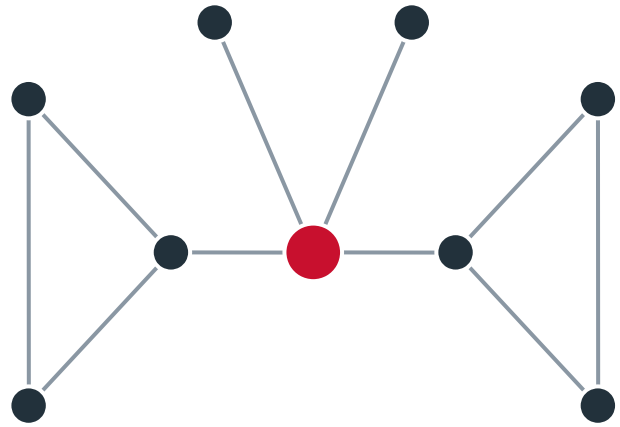
Cohesive

Everyone is within short reach; several routes; losing one tie changes little.



Fragmented

Three groups never communicate through the observed ties.



Brokered

Two groups reach each other only through one person, whose departure would split the organization.

Four mechanisms by which structure matters

Access

A tie is a channel for information, resources, introductions. A short path to an expert is an opportunity to learn, not what is actually passed on: say what is supposed to travel.

Brokerage

A contact across a gap is non-redundant: early and varied access, some control over what passes. Rival: able people end up in broker positions, so the position can be a symptom.

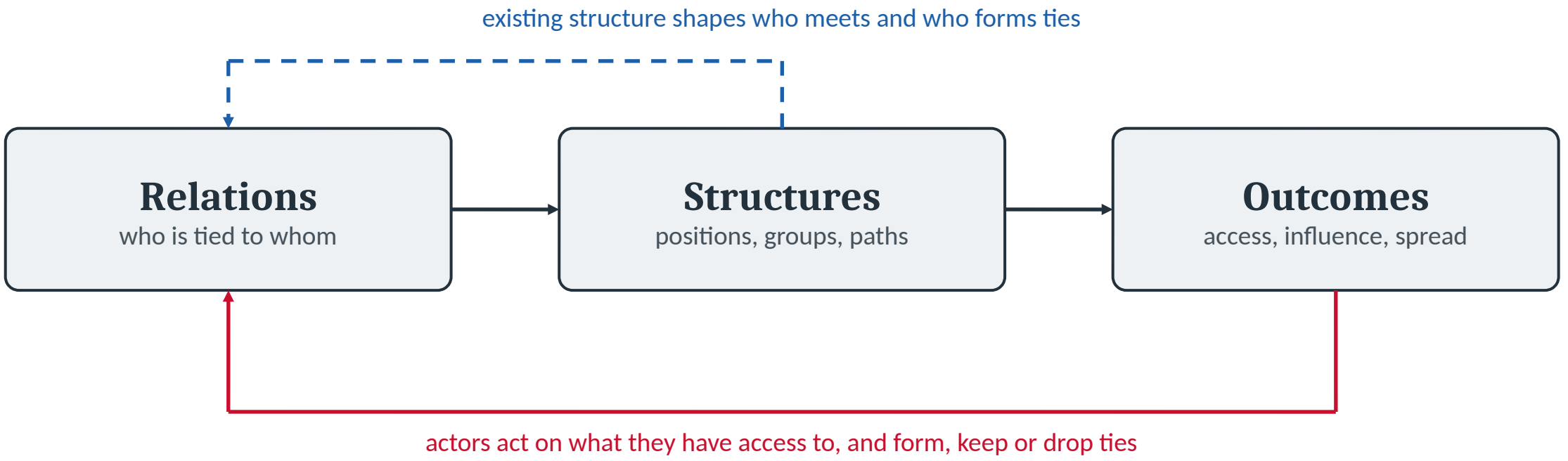
Closure

Dense local structure supports monitoring, reputation, trust. It creates the opportunity for monitoring; it does not show that anyone talks about whom.

Diffusion

The network supplies exposure; adoption depends on the actor. A bridge helps a rumor (one exposure suffices) and can hinder a practice that needs several.

Relations, structures, outcomes: a cycle, not a chain



Three questions organize the book: what patterns do relations form, what do the patterns do to those inside them, and how do the patterns arise and change? For the law firm we have one wave: the structure the cycle produced, not the cycle.

One association, three accounts: friends have similar grades

Influence or position

The tie or the position helps produce the outcome: friends study together, share attitudes to school. Something moves along the tie.

Selection or reverse causation

The arrow runs the other way: pupils with similar grades choose each other, and a successful pupil attracts requests for help.

Shared causes

Other factors affect both tie and outcome: the same class, the same teacher, the same neighborhood.

In the firm: partners are asked for advice more often than associates

Being consulted builds standing (position); competence made them partners and is what colleagues seek (selection); the firm assigns partners to lead cases (shared cause). Measurement can strengthen or hide any of the three: Chapter 3.

A network can be cause, outcome or context

Cause

Network properties explain an outcome. Does a broker earn more? Most positional measures of Part II serve this use.

Outcome

The ties themselves are explained. Why do two nodes become connected, why do triangles close? Models for tie formation (Part IV).

Context

The network is the medium of another process. An epidemic spreads differently through a clustered than a random contact network at identical density.

Use	The use fixes the problem
Cause	invites the rival accounts of the previous slide
Outcome	needs a model of tie formation that allows for dependence among ties
Context	is often best handled by simulation

Three questions for every measure, three mistakes

Whenever a measure appears, ask

1 What theoretical claim is it meant to represent?

2 What would have to be true of the data for it to represent that claim?

3 What else could produce the same result?

Having a command for a measure is never a reason to use it.

What can go wrong

- Equating exposure with adoption: the network supplies the exposure only.
- Inferring influence from a cross-sectional similarity, or from a group mean: the reflection problem.
- Calling any data set with several units a network: without data on pairs it is an ordinary data set.

Hands-on (10 min)

Ch. 1, Ex. 4

Look at the three networks of nine people on the earlier slide. Name the network in which (a) a rumor started by any one person would reach everybody in the fewest steps, (b) the loss of one person would do the most damage to communication, and (c) some people could never hear about each other's work at all. One sentence of reasoning for each.

Hint Follow the lines. In the brokered network, cover the large dot with a finger and see what remains; in the fragmented network, try to draw a route from one triangle to another.

Ch. 1, Ex. 5

Assign each study to one use of a network in an argument: as a cause, as an outcome or as a context. (a) A study asking whether central employees are promoted faster. (b) A study asking why some pairs of countries form trade agreements and others do not. (c) A study asking how quickly a rumor spreads through a company depending on its organizational chart.

Hint For each study, name what is being explained. Then ask where the network sits: among the explanations, in the thing explained, or as the route along which the explained process travels.

What to take away

Relations are properties of pairs

Structures built from them have properties no single tie or person has; many outcomes depend on those structures, not on attributes alone.

A measure is a description until it is tied to a claim

A network can enter an argument as a cause, an outcome or a context, and the same association can arise from influence, selection, shared causes or the way the ties were recorded.

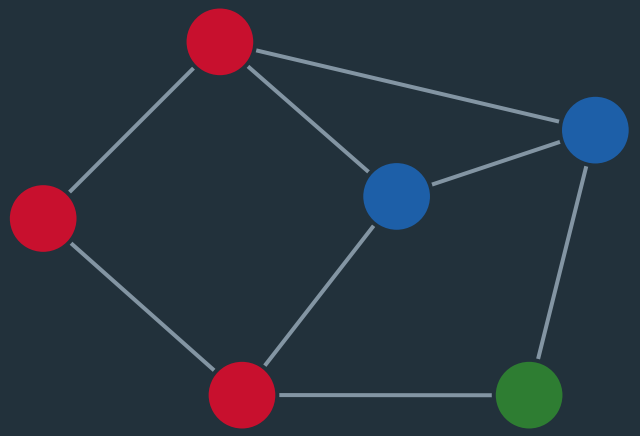
The method follows the question

Network analysis works best when the research question, a theoretical argument, suitable relational data and a method fit together.

2

What is a network?

55 min · Chapter 2 · four associates, then the lawyers

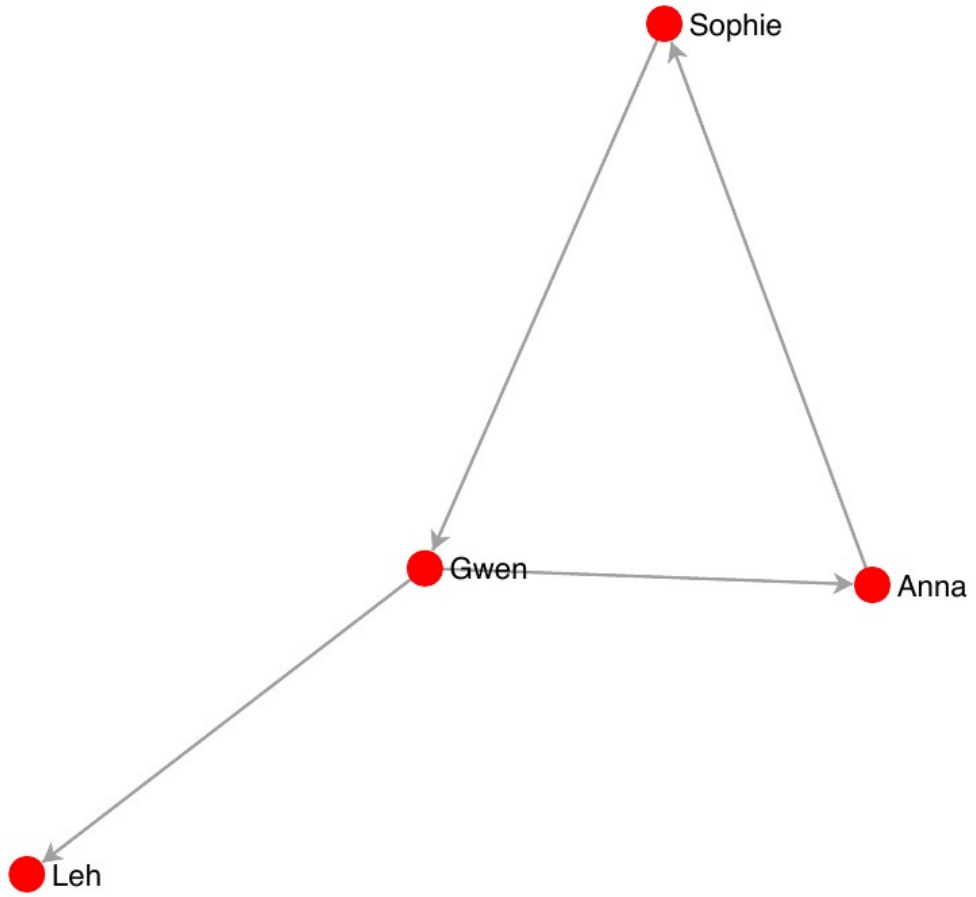


Lines or arrows? Both drawings are faithful to the data

- Anna names Sophie.
- Sophie names Gwen.
- Gwen names Anna and Leh.
- Leh names no one.

Draw lines and Leh is Gwen's friend. Keep the arrows and Leh is someone Gwen counts as a friend and who counts nobody.

Only the second drawing says that Leh names no one in return.



The four-person network (Chapter 2): no nomination is returned.

Every choice of type drops some information

Choice	Options	Recording it as the simpler type discards
Direction	directed (an arc) or undirected (an edge)	who initiated the tie; reciprocity is only defined for directed ties
Value	binary (present or absent) or valued (weighted)	strength: three joint papers against one, fifty e-mails against one
Sign	positive, or negative as well	hostility cannot be a 0: a -1 is a tie, a missing tie is not
Loops	a tie from a node to itself	allowed only where the relation has them (self-citation: yes; naming oneself as a friend: no)

Record a relation as directed whenever the data say who named whom; symmetrize only as a deliberate, stated step. Keep the values and dichotomize only for a measure that needs a binary input, stating the threshold.

Edge list and adjacency matrix: one network, two views

ego	alter
Anna	Sophie
Sophie	Gwen
Gwen	Anna
Gwen	Leh

Edge list: one row per tie; its length grows with the number of ties.

	Anna	Sophie	Gwen	Leh
Anna	0	1	0	0
Sophie	0	0	1	0
Gwen	1	0	0	1
Leh	0	0	0	0

Adjacency matrix: row = sender, column = receiver; its size grows with the number of nodes.

4 / 12

four ties among 4 × 3 possible ordered pairs: density 0.33

Two weaknesses

A person with no ties does not appear in an edge list. A matrix is wasteful for a large sparse network. Pick the form by the task.

Declare the network once: nwset, then nwsummarize

```
. input str8 ego str8 alter
  "Anna" "Sophie"
  "Sophie" "Gwen"
  "Gwen" "Anna"
  "Gwen" "Leh"
  end
. nwset ego alter, edgelist directed name(mini)
. nwsummarize mini
   Directed: true
   Nodes: 4
   Arcs: 4
   Density:  .333
   Mean outdegree:  1.00
```

nwset

Declares the two variables as an edge list and the network as directed. Typed once; stored; every later command already knows it.

nwsummarize

Reads directedness, node and arc counts and density from the object. Density .333 is the 4/12 computed by hand.

Eight of the 71 lawyers: small enough to check by hand

```
. nwwebuse lazega, nwclear
. nwset
(3 networks)
      lazega_adv
      lazega_cow
      lazega_fr
. list _nwnode seniority status office practice ///
>     if mod(seniority, 9) == 1, clean noobs
. nwsubset lazega_adv if mod(seniority, 9) == 1, ///
>     name(adv8)
```

_nwnode	seniority	status	office	practice
n1	1	partner	Boston	litigation
n10	10	partner	Boston	corporate
n19	19	partner	Boston	corporate
n28	28	partner	Hartford	corporate
n37	37	associate	Providence	corporate
n46	46	associate	Hartford	corporate
n55	55	associate	Boston	litigation
n64	64	associate	Boston	corporate

Every ninth lawyer by seniority, starting from the top: four partners and four associates, five in Boston, two in Hartford, one in Providence. The number in the label is the seniority rank.

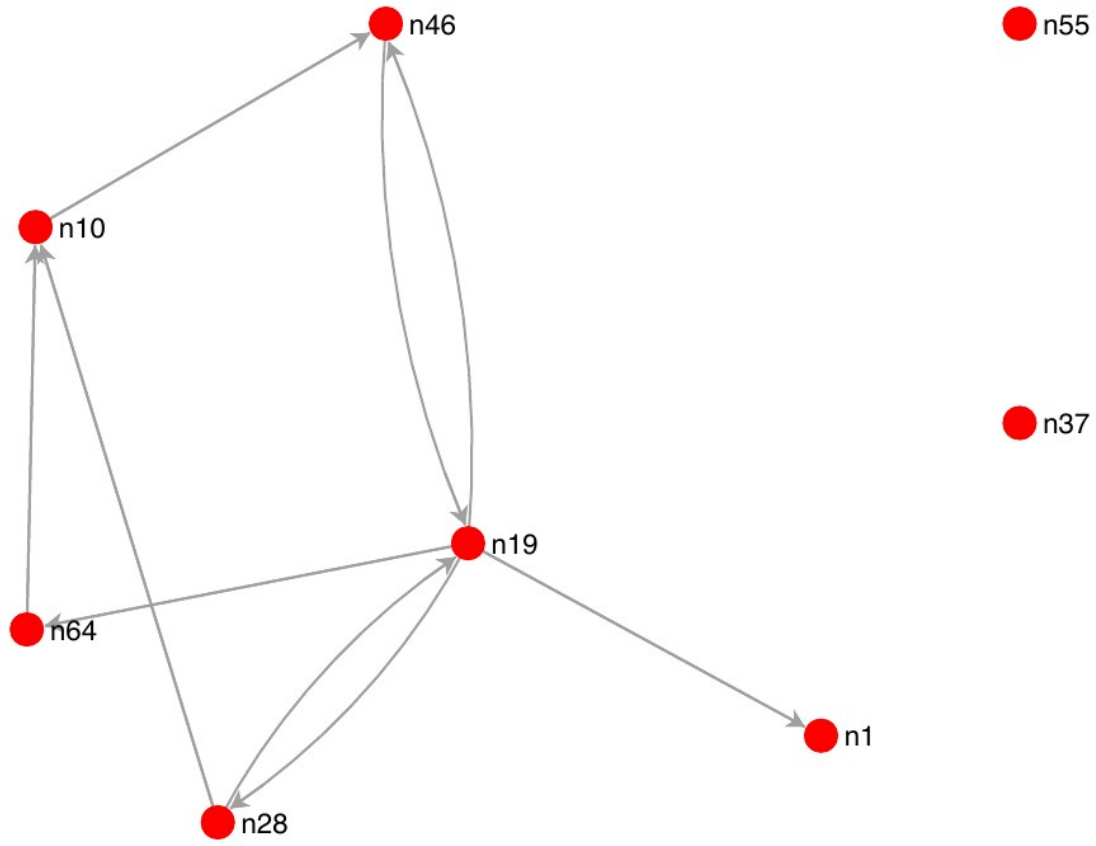
The whole firm: 71 lawyers, 36 partners (ranks 1 to 36), 35 associates, 3 offices.

adv8: nine arcs, density .161, and two lawyers without any tie

```
. nwsummarize adv8
   Nodes: 8
   Arcs: 9
   Density:  .161
   Mean outdegree:  1.12
. set seed 42
. nwplot adv8, label(_nwnode)
```

9 / (8 × 7) = 9 / 56 = .161

n37 and n55 have no advice tie to any of the other seven. In the full firm neither is isolated: only the set of people counted changed, and with it their position.



Advice among eight lawyers; an arrow runs from the lawyer who asks to the advisor.

The edge list forgets the isolates; the matrix keeps everyone

```
. quietly nwtoedge adv8, ego(sender) alter(receiver)
. list sender receiver if adv8 == 1, clean noobs
    sender   receiver
       n10        n46
       n19         n1
       n19        n28
       n19        n46
       n19        n64
       n28        n10
       n28        n19
       n46        n19
       n64        n10
```

```
. nwsummarize adv8, matonly
        1   2   3   4   5   6   7   8
  +-------------------------------------+
1 |   .   0   0   0   0   0   0   0   |
2 |   0   .   0   0   0   1   0   0   |
3 |   1   0   .   1   0   1   0   1   |
4 |   0   1   1   .   0   0   0   0   |
5 |   0   0   0   0   .   0   0   0   |
6 |   0   0   1   0   0   .   0   0   |
7 |   0   0   0   0   0   0   .   0   |
8 |   0   1   0   0   0   0   0   .   |
  +-------------------------------------+
```

n37 and n55 appear in no row of the list, but own the all-zero rows and columns 5 and 7 of the matrix. The diagonal is missing because a lawyer cannot ask himself or herself. Rows are senders (who asks), columns receivers; a source with the opposite convention flips every arrow without an error message.

All 71 lawyers: density by hand, and the spread of degrees

```
. nwwebuse lazega, nwclear
. nwsummarize lazega_adv
   Nodes: 71
   Arcs: 892
   Density:  .179
   Mean outdegree:  12.56
. display 892 / (71 * 70)
.17947686
. quietly nwdegree lazega_adv, generate(adv_out adv_in)
. summarize adv_out adv_in
    Variable |   Obs       Mean   Std. dev.   Min   Max
     adv_out |    71   12.56338    8.294116     2    37
      adv_in |    71   12.56338    5.615622     3    26
```

.179

= 892 / 4,970 ordered pairs

2 to 37

outdegree (asking), against 3 to 26 for indegree (being asked)

Mean outdegree and mean indegree are both 12.56: an identity of every directed network, not a finding.

Symmetrizing costs information: 892 arcs become 700 edges

```
. nwsym lazega_adv, generate(adv_und)
. nwsummarize adv_und
   Directed: false
   Nodes: 71
   Edges: 700
   Density:   .282
   Mean degree:  19.72
```

700

edges, down from 892 arcs: the 192 mutual pairs lose one tie each

.282

density of the undirected network (÷ 2,485), against .179 for the directed one (÷ 4,970)

It says

In 700 of the 2,485 pairs of lawyers at least one names the other: 192 name each other and in 508 only one does.

It does not say

That the lawyers are more connected than a moment ago: .282 divides by 2,485 unordered pairs and .179 by 4,970 ordered pairs. Nor who went to whom for advice, which the status argument of Chapter 1 needs.

Hands-on (15 min)

Ch. 2, Ex. 3

Load the lawyers with nwwebuse lazega, nwclear and run nwsummarize lazega_cow, the work network. Compute its density by hand from the number of arcs and compare it with the printed value. Then symmetrize it with nwsym lazega_cow, generate(cow_und), run nwsummarize cow_und, and compute the density of the undirected network by hand as well. Why is the undirected density higher, although no lawyer gained a tie?

Hint The directed network divides by 71 · 70 = 4,970 ordered pairs, the undirected one by 71 · 70 / 2 = 2,485 unordered pairs. Take one pair with an arc in only one direction: how much does it add to the count of ties and to the number of pairs before and after symmetrizing?

Ch. 2, Ex. 7

Do for the friendship network what we did for advice: run nwsummarize lazega_fr, then nwsym lazega_fr, generate(fr_und) and nwsummarize fr_und. From the number of arcs and the number of edges alone, work out how many pairs name each other as friends and in how many only one names the other. Do the same for advice (892 arcs, 700 edges). In which relation is a larger share of the tied pairs mutual?

Hint A mutual pair holds two arcs but becomes one edge; a one-sided pair holds one arc and becomes one edge. So arcs minus edges gives the number of mutual pairs, and edges minus mutual pairs the number of one-sided pairs.

What to take away

A tie is defined by a choice of type

Directed or undirected, binary or valued, signed or not: each choice keeps some information and discards other information.

Edge list and matrix are exact and interchangeable; conversions are not

Symmetrizing, dichotomizing and restricting to a subset each change the network, and each should be made on purpose and stated.

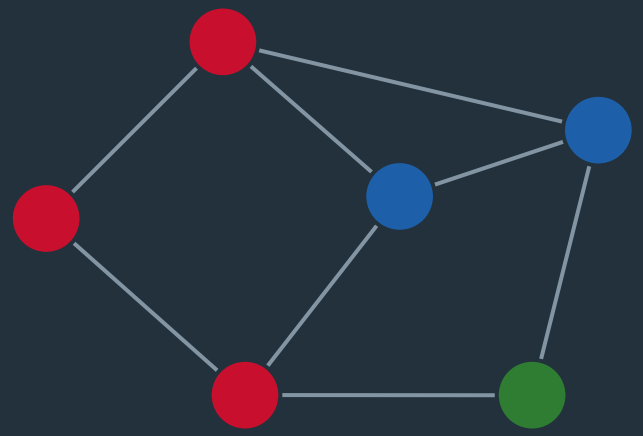
Density is an average and says nothing about the spread

Report it with the degrees behind it: .179 for advice hides outdegrees from 2 to 37. A subset is not the network: .161 and two isolates among eight lawyers, .179 and none among 71.

Break

15 minutes

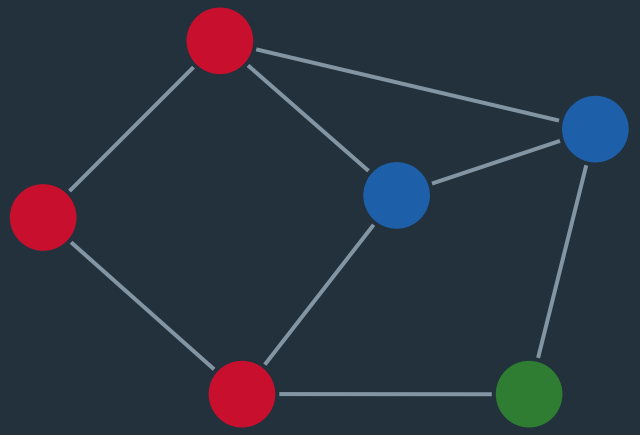
Back at 11:00



3

From a research question to a network and a method

50 min · Chapter 3 · the reporting checklist



There is no 'the' friendship network of a school class

Researcher A

Each pupil names up to three friends. A tie is kept only for pairs who named each other.

Researcher B

A class list with a yes or no for every name. A tie is kept for every pair named by at least one.

Nobody made an error, and neither network is false.

How many friends a pupil may name, whether a one-sided nomination counts, who is inside the school, how often the question is asked: each choice produces a different network. The instrument shapes the result, and a network is an operationalization of a social fact, like a survey item for an attitude.

Six decisions turn a question into a network

Boundary

Realist (the actors' own sense of who belongs) or nominalist (the researcher's criterion)?

Tie

Which relation exactly: affective bond or recent interaction? Directed? Valued?

Instrument

Name generator, with or without a ceiling on names, or a roster of everybody?

Design and reports

Whole network or ego network? One-sided reports: AND, OR, or kept directed?

Missing data

Response rate; missing ties dropped, imputed or treated as structural zeros?

Time

One wave, repeated waves, a continuous record?

The running question: how is informal advice-seeking organized within the firm, and does it follow the formal hierarchy or cut across it?

The boundary decides which measures mean anything

Realist boundary

Follows the actors' own sense of who belongs: a club, a gang. Fails where members disagree or the group is one nobody has named.

Nominalist boundary

Imposed by the researcher: a membership list, authors of a set of papers. Fails where the criterion cuts across the relation.

- Distance and betweenness are computed inside the boundary and say nothing about paths that leave it.
- A firm's roster is reasonable for reporting lines and a poor boundary for informal advice, which routinely leaves the building.
- Choose the boundary from the relation being studied, not from administrative convenience, and say what it leaves out.

Name generator or roster: the instrument shapes the degrees

	Name generator	Roster
Asks	'Who do you turn to for advice?' Respondent supplies names, up to a ceiling or without one	Full list of possible alters; a relation to each
Strength	Scales to large, unbounded populations	Reduces recall problems; nobody forgets a colleague on the list
Weakness	Recall suffers past a handful of names; wording matters	Burden grows with population; late names tend to get 'no' (fatigue)

Randomize the order of a roster across respondents, so that fatigue hits different names for different respondents.

6

friends per pupil allowed in the Glasgow study

12

classmates allowed in the klas12b study

A pupil with exactly six nominations might have exactly six friends, or more. Report the ceiling and how many respondents reached it.

Whole or ego network, and the shape of missing data

	Whole network	Ego network
Collects	ties among every member of a bounded population	each respondent's alters, at most the ties among them
Reaches	bounded groups: firm, class, gang	large or loosely bounded populations: a national sample
Cannot give	ego-network scale at low cost	closeness, betweenness, communities (Chapters 8, 9)
Survives	everything of Part II	ego size, density, composition, effective size, constraint (Chapter 11)

Missing data is not a missing cell

A missing respondent removes every tie they would have reported. A class with 80% response is not a smaller version of the full network.

A response rate is not a footnote

Report it, say how missing ties were handled, and check whether the main results change when non-respondents are dropped.

The reporting checklist: six questions

	Topic	The question
1	Boundary	What defines population membership: a positional rule (an organization), an event-based rule or a reputational rule, or a combination?
2	Tie	What relation, exactly, does a tie represent, and how was 'present versus absent' (or the level of a valued tie) operationalized in the instrument?
3	Instrument	Name generator or roster? If a name generator: a stated nomination ceiling, and how many respondents actually reached it?
4	Design and reports	Whole or ego network? If whole: which convention resolved disagreement between ego and alter (AND, OR, or kept directed)?
5	Response and missing	What was the response rate, and how were non-responding units' missing ties handled (dropped, imputed, structural zeros)?
6	Time	What is the temporal boundary (one wave, or intervals), and does the research question need more than one wave?

The checklist applied to the lawyers' data

	Question	The lawyers' data
1	Boundary	All members of one firm, three offices, partners and associates: a positional boundary. Outside advisors are not in the network.
2	Relation	Three directed relations: advice, work together, friendship. Binary.
3	Instrument	A roster. No low ceiling is evident (outdegree 2 to 37); a formal ceiling would have to come from the documentation.
4	Design, reports	Whole network. Directed ties kept as reported; OR and AND give 501 and 74 friendship edges.
5	Response, missing	Every lawyer has outgoing advice nominations, which is not the same as participation. The response rate is not given with the data.
6	Time	One wave.

Reading the boundary off the data

```
. nwwebuse lazega, nwclear
. tab office status
           |        status
    office | associate    partner |     Total
-----------+----------------------+----------
    Boston |        26         22 |        48
  Hartford |         6         13 |        19
Providence |         3          1 |         4
-----------+----------------------+----------
     Total |        35         36 |        71
```

What the file shows

71 nodes: 48 in Boston, 19 in Hartford, 4 in Providence; 35 associates and 36 partners. Office and status are not independent.

What it cannot show

The lawyers had advisors outside the firm. Every distance computed later is computed inside it: a limit to state, not to hide.

Whose report counts: 501 friendship ties or 74?

```
. nwdyads lazega_fr
    Mutual |      Asym |      Null
        74 |       427 |      1984
    Reciprocity: .148
. nwsym lazega_fr, generate(fr_or) mode(max)
. nwsym lazega_fr, generate(fr_and) mode(min)
```

501

OR rule, mode(max): density .202

74

AND rule, mode(min): density .030

575

arcs, directed: density .116

It says

The same survey answers give about 6.8 times as many ties under the OR rule as under the AND rule. In the 74 pairs both lawyers agree; the other 427 pairs rest on one report only.

It does not say

Which network is the true friendship network. The AND rule is not cleaner for being stricter, since it discards every one-sided report, and the OR rule is not more complete for being larger.

From network to method: describe, test or model

Describe

What does the network look like? Chapters 5 to 14: pictures, density and census, paths, centrality, groups, positions, ego networks, change.

Compare or test

Is a pattern unusual, or associated with something else? Baselines and simulation (Part III), then Chapters 18 to 20.

Model

Which dependencies and processes could account for ties and change? Chapters 21 to 28.

- The data structure says what is possible: one observed network up to Chapter 23; panel waves for Chapters 24 to 26; timestamped events for Chapters 27 and 28.
- A centrality score is evidence of a position, not of influence; a significant coefficient is evidence against a stated null, not of a cause.

Hands-on (15 min)

Ch. 3, Ex. 2

Five respondents report their friendships. For 6 pairs of them, both people in the pair name each other; for 4 further pairs, one person names the other but is not named back; no other pair is named. How many undirected ties does the AND rule produce, how many does the OR rule produce, and how many arcs does the directed network have?

Hint The AND rule keeps only the pairs in which both name each other; the OR rule keeps those and the one-sided pairs. In the directed network a pair in which both name each other holds two arcs.

Ch. 3, Ex. 3

The chapter applied the OR and AND rules to friendship and found 501 against 74 edges. Do the same for advice: after nwwebuse lazega, nwclear, run nwdyads lazega_adv, predict the edges under each rule from the dyad census, then check with nwsym lazega_adv, generate(adv_or) mode(max), nwsym lazega_adv, generate(adv_and) mode(min) and nwsummarize adv_or adv_and. Report both edge counts, densities and their ratio. Does the rule matter more for advice or for friendship?

Hint The AND rule keeps the mutual pairs; the OR rule keeps the mutual and the asymmetric pairs. Both densities divide by the $71 \cdot 70 / 2 = 2{,}485$ unordered pairs.

Your own data: fill in the checklist

	Question	Your data set
1	Boundary	
2	Relation	
3	Instrument, ceiling	
4	Whole or ego; reconciliation of reports	
5	Response rate and missing ties	
6	Temporal boundary	

Ch. 3, Ex. 10

Choose a data set from your own field, or one of the book's, and answer the six questions for it. Which answer was hardest to find? Which choice would you change if the question were about information diffusion rather than emotional support?

What to take away

A network is constructed, not found

Boundary, tie definition, design, missing data and time window each shape what it can answer, and each should be decided deliberately and reported.

The rule that turns reports into ties can change the number of ties several times over

Friendship gives 501 ties under the OR rule and 74 under the AND rule; the rule belongs in every description of the data.

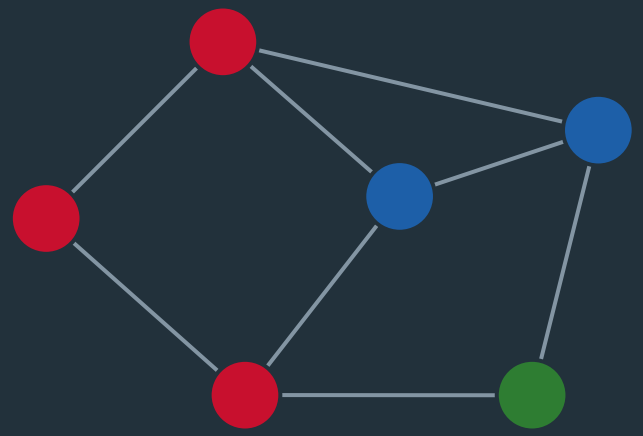
The choice of method is the second half of the same design

The question, the data structure and the goal (describe, compare or test, model) together point to a family of methods.

Lunch

60 minutes

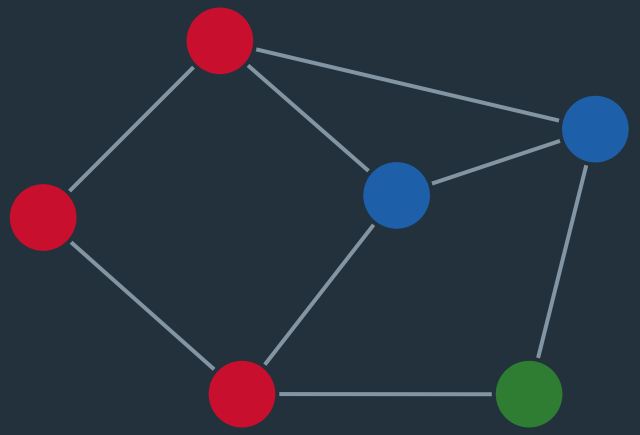
Back at 12:50



4

Working with networks in Stata

85 min · Chapter 4 · mostly at the computer



Networks sit beside the data and are found by name

_nwnode	seniority	status	office	practice
n1	1	partner	Boston	litigation
n10	10	partner	Boston	corporate
n11	11	partner	Boston	litigation
n12	12	partner	Boston	corporate
...	...	...	...	...

The data set: 71 lawyers, one row each, 13 variables; attributes are ordinary variables.

lazega_adv
advice · 892 arcs · density .179

lazega_cow
work together · 1,104 arcs · density .222

lazega_fr
friendship · 575 arcs · density .116

Network objects: not variables, stored beside the data, found by name (nwset lists them).

Every command finds a network by name; every result that belongs to a node comes back as a variable. **So tabulate, regress and the rest of Stata work on network measures directly.**

Declare from a matrix, or an edge list with isolate rows

```
. nwclear
. nwset, mat((0,1,0,0\0,0,1,0\1,0,0,1\0,0,0,0)) ///
>         directed name(toy) labs(Anna, Sophie, Gwen, Leh)
. nwdegree toy, generate(_outdegree _indegree)
(output omitted)
```

The same four arcs as in the edge list of the morning: a network does not change when it is written down in another format.

Row i, column j holds the arc from i to j; rows are separated by a backslash; labs() names the nodes in row order.

The matrix keeps every actor

Every actor owns a row, so isolates are safe. The limit is size: a cell for every pair.

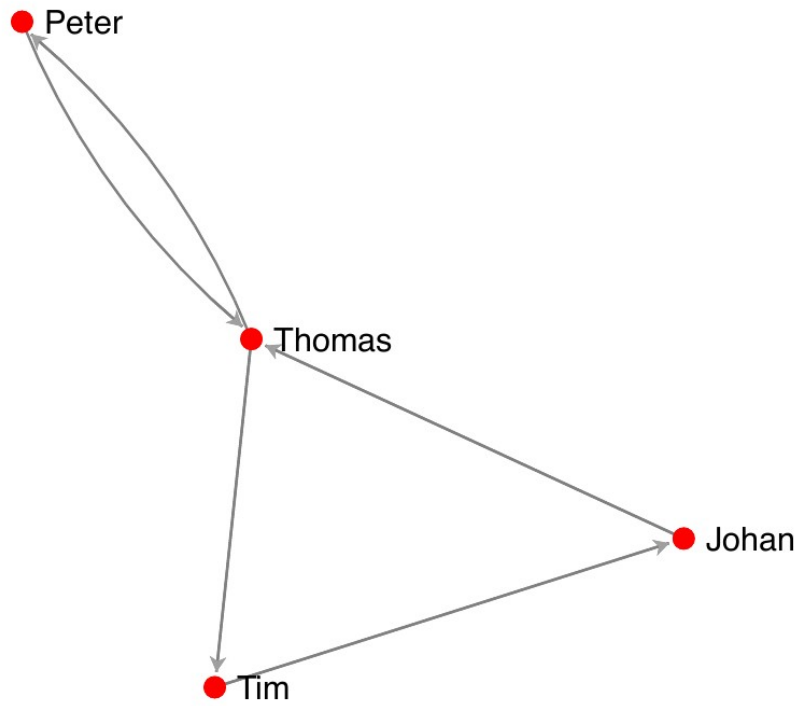
The edge list drops isolates

A person nobody names and who names nobody vanishes: n shrinks, density rises. Remedy: a row with a sender and an empty receiver.

Import: nwimport reads Pajek, UCINET, GML and edge-list files

```
. nwclear
. nwimport
"https://nwcommands.org/examples/edgelist_example.txt", ///
>         type(edgelist) name(tiny)
------------------------------
Importing successful
. nwsummarize tiny
   Directed: true
   Nodes: 4
   Arcs: 5
   Density:  .417
```

Inference is a convenience, not a guarantee. Here nwimport infers directedness correctly (Thomas to Peter and Peter to Thomas are separate rows). Check the Directed and Valued lines of nwsummarize against what you know about the relation before going on.



Four people, five arcs, read from a two-column file.

A survey export: reshape long, then an edge list

```
. input str6 id str6 friend1 str6 friend2 str6 friend3
  "anna"   "sophie" "gwen"   ""
  "sophie" "anna"   ""       ""
  "gwen"   "anna"   "leh"    "fion"
  "leh"    ""       ""       ""
  "fion"   "gwen"   "sophie" ""
  end
. reshape long friend, i(id) j(nomination)
. drop if missing(friend) & nomination > 1
(6 observations deleted)
```

5 → 15 → 9

respondents, rows after reshape long, rows after dropping empty slots

- **nomination > 1 keeps Leh**

  Leh names nobody, so his first slot stays as a row with an empty receiver. A plain drop if missing(friend) would delete all three of his rows.

From nominations to ties: undirected is the union rule

```
. nwfromedge id friend, name(friends) undirected
. nwsummarize friends
   Directed: false
   Nodes: 5
   Edges: 5
   Density:  .5
```

8 → 5

nominations become undirected ties under the union rule (either person reports it)

3

ties under the intersection rule: only ties both persons report

Tie	Named from
Anna-Sophie	both ends
Anna-Gwen	both ends
Gwen-Fion	both ends
Gwen-Leh	one end
Fion-Sophie	one end

Density .5 is 5 of 10 possible ties; under the intersection rule it would be 3 of 10 = .3. The rule changes density before a single substantive measure is computed.

Hands-on (20 min)

Ch. 4, Ex. 1

Declare a five-person network from an adjacency matrix with nwset, mat(): Anna names Sophie and Magalie, Sophie names Anna, Gwen names Leh, Leh names Gwen and Magalie, and Magalie names no one. Name it mini5. Before running nwsummarize, work out the nodes, arcs and density by hand, then check. Which pairs are reciprocated?

Hint Row i, column j holds a 1 for an arc from i to j; rows are separated by a backslash; labs() gives the names in row order. Add directed. The density divides the arcs by n(n-1). A pair is reciprocated when both its cells hold a 1.

Ch. 4, Ex. 2

In the survey example, suppose Gwen had not named Leh, so that nobody names Leh and Leh names nobody. Rebuild the undirected network twice from the long data: once with drop if missing(friend) & nomination > 1, once with a plain drop if missing(friend). Report nodes, edges and density for each, and say which version is right and why.

Hint After reshape long, remove Gwen's nomination of Leh with drop if id == "gwen" & friend == "leh", then nwfromedge id friend, name(name) undirected after each drop rule. Use preserve and restore so both versions start from the same data.

Several networks, one data set: use wildcards

```
. nwwebuse lazega, nwclear
. nwsummarize lazega_*
(prints advice, work and friendship in turn)
. quietly nwdegree *_adv, generate(adv_out adv_in)
. summarize adv_out adv_in
  Variable |  Obs      Mean   Std. dev.  Min  Max
   adv_out |   71  12.56338  8.294116    2   37
    adv_in |   71  12.56338  5.615622    3   26
```

Pattern	Matches
`lazega_adv`	that network only
`lazega_*`	all three lawyers' networks
`*_adv`	the advice network
`*_fr`	the friendship network
`glasgow?`	glasgow1, glasgow2, glasgow3: the three waves

`*` stands for any string of characters, `?` for exactly one. Choose names that sort and match well: a common prefix for one data source, a common suffix for one relation. The name is the only handle you have.

892 advice, 1,104 work and 575 friendship arcs: three relations among the same 71 lawyers.

Between sessions: nwsave, nwuse, and nwload

```
. nwclear
. set seed 12345
. nwrandom 20, ntimes(5) prob(.2)
. nwsave mynets, replace
. nwclear
. nwuse mynets, nwclear
. nwset
(5 networks)
      random_1 ... random_5
```

The file holds the networks, not the commands or the seed that produced them: the do-file remains the record of how they came about.

nwsave and nwuse

Mirror Stata's save and use. A .nwdta file holds the networks together with the variables in memory; Stata's own save would lose the networks.

nwload: opt-in

Writes a network as one variable per node: 18 variables for the 16 Florentine families, 71 more for the lawyers. For writing a new command or handing the network to a non-network procedure.

nwname: what was declared, not whether it is right

```
. nwwebuse lazega, nwclear
. nwname lazega_adv
. return list
scalars:
             r(nodes) =  71
         r(selfloops) =  0
      r(missing_edges) =  71
macros:
            r(valued) : "false"
          r(directed) : "true"
              r(labs) : "n1,n10,n11,n12,n13,n14,.."
```

Check the declaration

directed is true because nwset was told so; checking it against the design is still up to you. Every other command reads this silently.

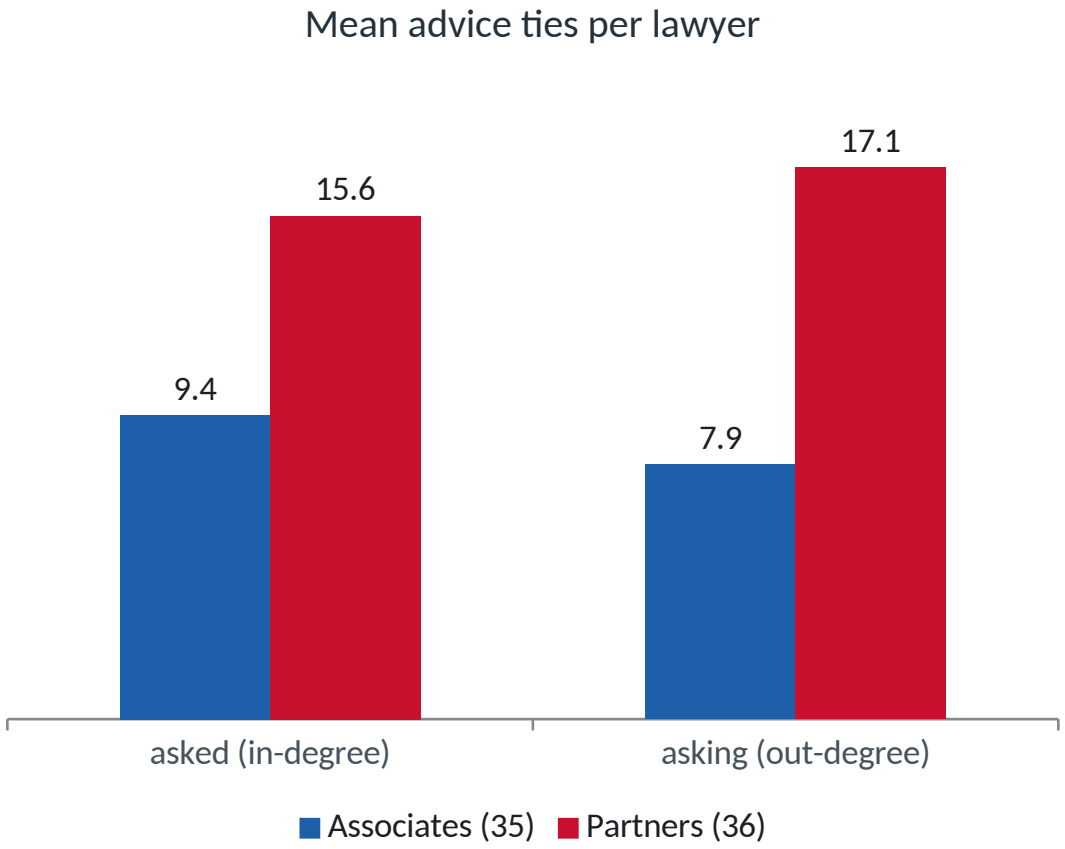
Nodes are stored alphabetically

n1, n10, n11, … not n1, n2, n3. 'Node 2' is n10. Commands that take a node by position count in this order.

Node-level results are variables, so Stata does the rest

```
. nwdegree lazega_adv, generate(adv_out adv_in) silent
. tabstat adv_in adv_out, by(status) ///
>         statistics(mean sd n)
  status |    adv_in   adv_out
---------+--------------------
associate |  9.428571  7.885714
  partner | 15.61111  17.11111
    Total | 12.56338  12.56338
```

Partners are asked more (15.6 against 9.4) and ask more (17.1 against 7.9). Both degrees are variables now: tabstat, ttest, regress, graph work on them like on any other variable. What the difference means is the rival-accounts question of Chapter 1.



Build new networks from old: nwgenerate, nwreplace

```
. capture noisily nwgenerate lazega_fr = lazega_adv * lazega_fr
Network lazega_fr already exists. Change netname or specify option replace.
. nwgenerate both = lazega_adv * lazega_fr
. nwsummarize both
   Arcs: 358
   Density:  .072
. nwreplace both[2,1] = 0
```

358

arcs along which a lawyer both asks a colleague for advice and names that colleague a friend

62% of the 575 friendship arcs, but only 40% of the 892 advice arcs. It says nothing about the direction of influence; how much two unrelated relations of these densities would overlap is the benchmark of Chapter 13.

As for variables

generate creates something that must not exist yet, replace overwrites something that must. The failed command would have overwritten the friendship network.

Vary the decision: subset, symmetrize, transpose, copy

```
. nwsubset lazega_adv if office == "Boston", name(adv_boston)
. nwsummarize adv_boston
   Nodes: 48
   Arcs: 539
   Density:  .239
. nwkeepnodes lazega_adv, nodes(1/10) generate(adv_first10)
. nwsymmetrize lazega_fr, mode(max) generate(fr_or)
. nwsymmetrize lazega_fr, mode(min) generate(fr_and)
. nwduplicate lazega_adv, name(adv_copy)
```

The originals stay untouched, so the same analysis can run on several versions and report how much the answer depends on the choice.

Command	Derives
`nwsubset ... if`	nodes that satisfy a condition on the attributes
`nwkeepnodes`	nodes named by number or label
`nwsymmetrize (nwsym)`	mode(max) OR rule, mode(min) AND rule
`nwtranspose`	reverses every arc: who is asked by whom
`nwduplicate`	an exact copy

Boston keeps 48 lawyers and 539 of 892 arcs: density .239 against .179, because the pairs across offices are no longer counted.

Hands-on (25 min)

Ch. 4, Ex. 4

After nwwebuse lazega, nwclear, write one nwsummarize command with a wildcard that reports all three networks without naming them, and one nwdegree command with a suffix pattern for friendship only, storing fr_out and fr_in. Summarize them and compare with the advice degrees. Then nwsave, nwclear, nwuse, and confirm with nwname lazega_fr and return list that r(directed) is still true.

Hint lazega_* for all three, *_adv for advice; the friendship name ends in _fr. nwdegree needs generate(fr_out fr_in). nwsave lawyers, replace and nwuse lawyers, nwclear read it back.

Ch. 4, Ex. 5

Build the network of arcs that are both work ties and friendship ties, lazega_cow * lazega_fr, with nwgenerate, and report how many of the 575 friendship arcs are also work arcs. Is that share larger or smaller than the share of friendship arcs that are also advice arcs (358 of 575)?

Hint Give the new network a name that does not exist yet, for example nwgenerate cowfr = lazega_cow * lazega_fr. The count is on the Arcs line of nwsummarize cowfr; divide it by 575.

What to take away

Declare a network once; every later command finds it by name and knows its type

The declaration of directedness and the rule that turns nominations into ties are claims about the relation, and they change density before anything else is computed.

Which command for which data

A list of ties: nwset ego alter, edgelist (directed or undirected, a row with an empty alter keeps an isolate). A matrix: nwset, mat() with labs(). A file: nwimport. A survey export: reshape long, then nwfromedge. Between sessions: nwsave and nwuse.

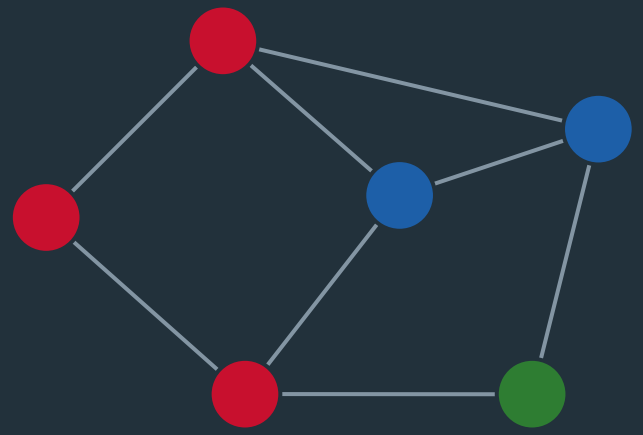
Materializing a network as plain variables is opt-in

nwload adds one variable per node; most commands never need it. Node-level results come back as ordinary variables, and r() holds the whole-network results.

Break

15 minutes

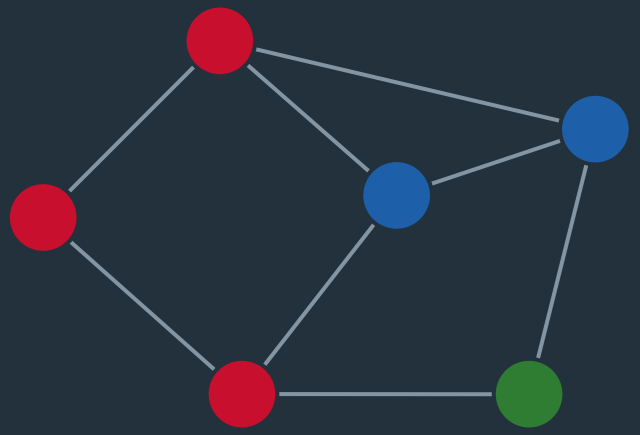
Back at 14:30



5

Seeing networks

75 min · Chapter 5 · the interactive pages

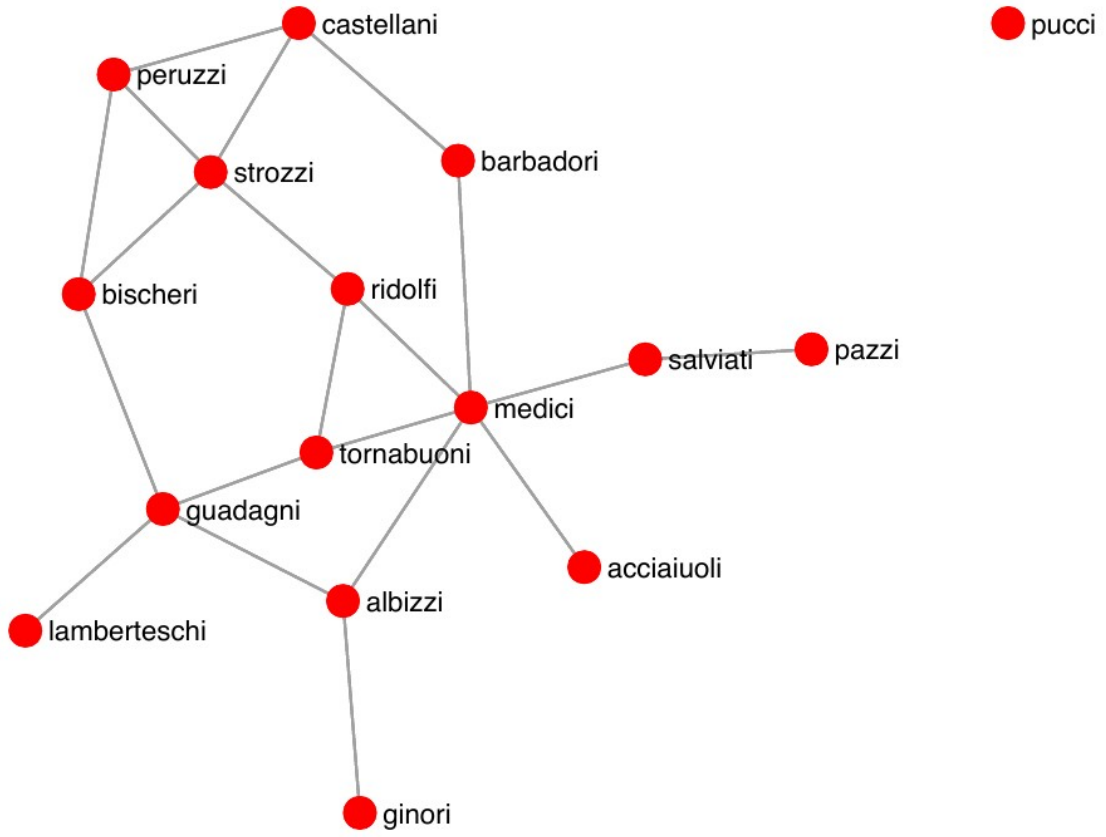


A first picture: the facts that survive any layout

```
. nwwebuse florentine, nwclear
. set seed 42
. nwplot flomarriage, label(_nwnode)
Calculating node coordinates...
Plotting network...
```

- Twenty marriages connect fifteen of the sixteen families.
- Pucci has no marriage tie to any other family.
- Medici, near the middle, has six ties.

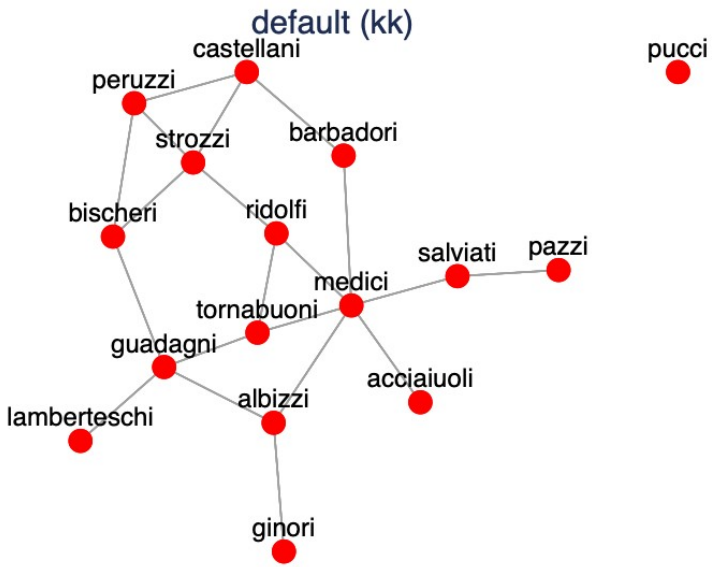
These facts are read from the lines and are true in any layout. The layout starts from random positions, hence the seed.

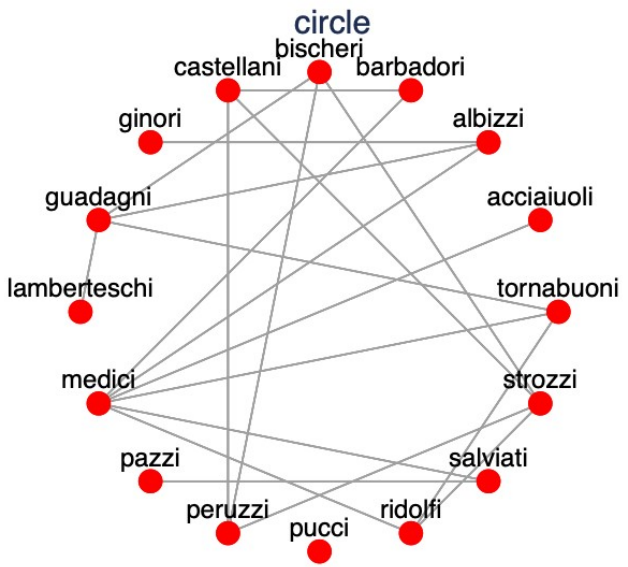


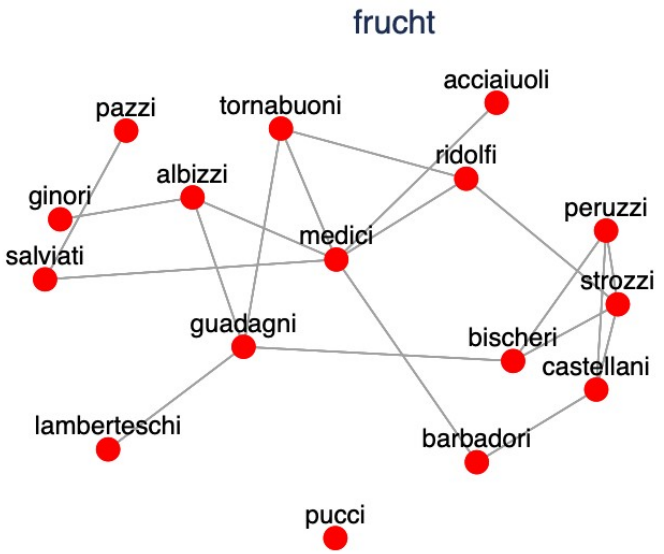
Marriage ties among sixteen Florentine families around 1430 (Padgett).

One network, three layouts, three stories

```
. nwplot flomarriage, label(_nwnode)                      // default kk, after set seed 42
. nwplot flomarriage, label(_nwnode) layout(circle)
. nwplot flomarriage, label(_nwnode) layout(frucht)   // after set seed 42
```







The ties are data; the positions, in most layouts, are not

Part of the picture	Data?	Why
Lines and arrows	Yes	They are the ties
Color, size, symbol mapped to a variable	Yes	A measured attribute: color(office), size(wealth)
Labels	Yes	A variable written next to each node
Position in kk, mds, frucht, circle, grid	No	An algorithm chose it so the picture is readable; the coordinates mean nothing about the actors
Position with real coordinates (map)	Yes	Longitude and latitude mean what they say, whatever the ties

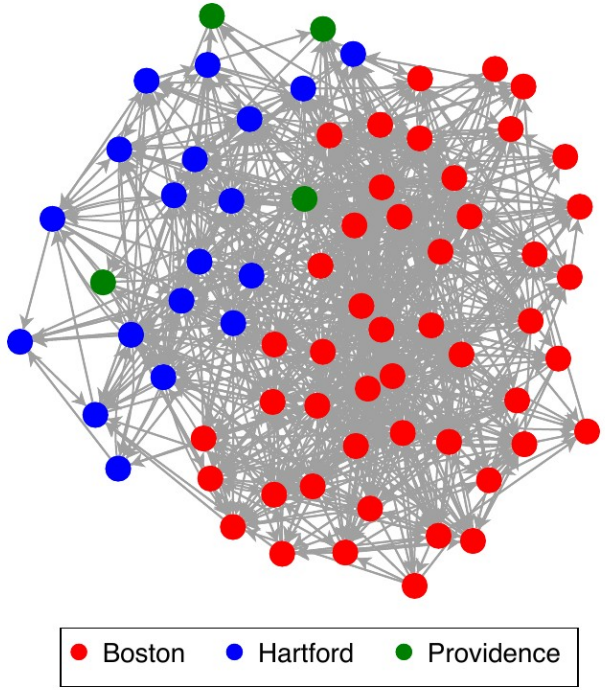
47.5

Medici's betweenness in every layout

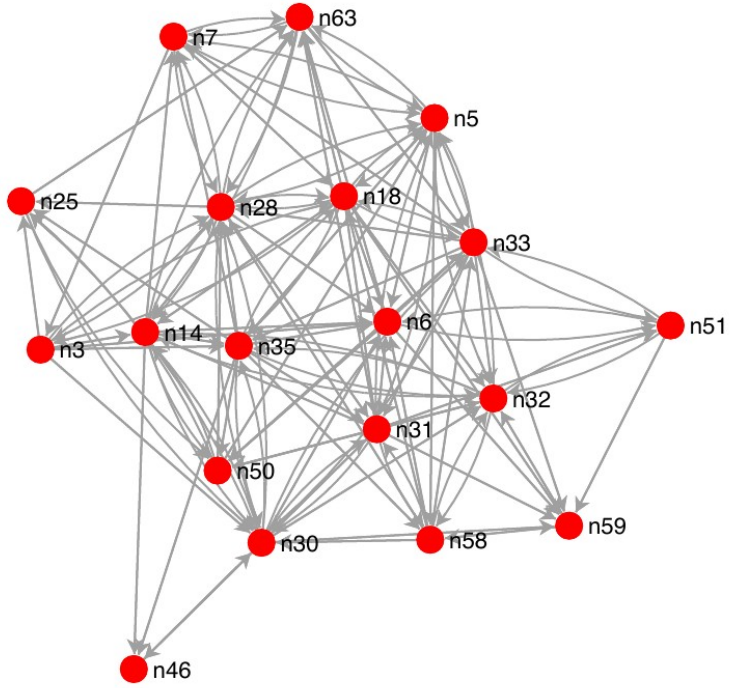
Reading a layout invites circularity. If the algorithm puts highly connected nodes in the middle, concluding from the middle that a node is important adds nothing. Claims about who is central belong to the measures of Chapter 8. A picture can illustrate a claim that a measure has established; it cannot establish one.

Attributes on the picture; draw less in a hairball

```
. nwplot lazega_adv, color(office)
. nwsubset lazega_adv if office == "Hartford", name(hartford)
. nwplot hartford, label(_nwnode)
```



71 lawyers, 892 ties: a hairball. Color is the office, a measured variable.

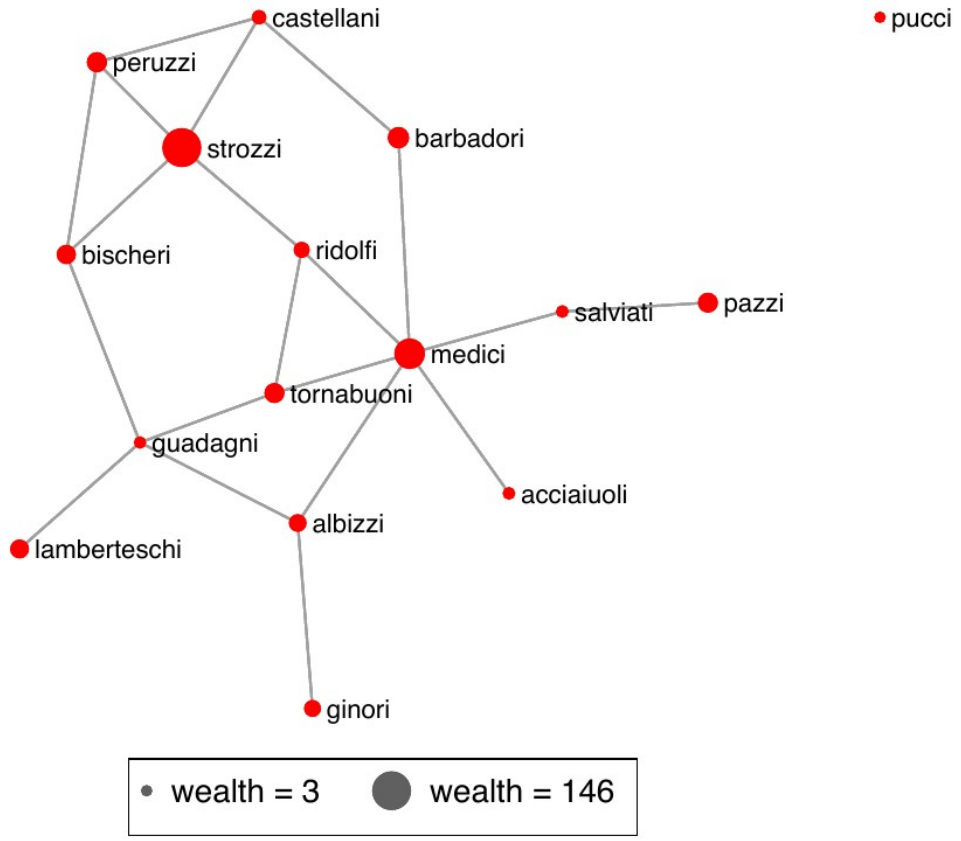


19 lawyers of one office: single ties can be read.

Size needs a quantity; a rank draws upside down

```
. set seed 42
. nwplot flomarriage, label(_nwnode) ///
>         size(wealth)
```

- Medici (103) and Strozzi (146) are the largest dots; Medici has six ties, Strozzi four.
- Whether wealth and marriage ties go together is a correlation to compute, not a reading of dot sizes.
- The lawyers' seniority is 1 for the most senior: as a size it draws the most senior smallest.



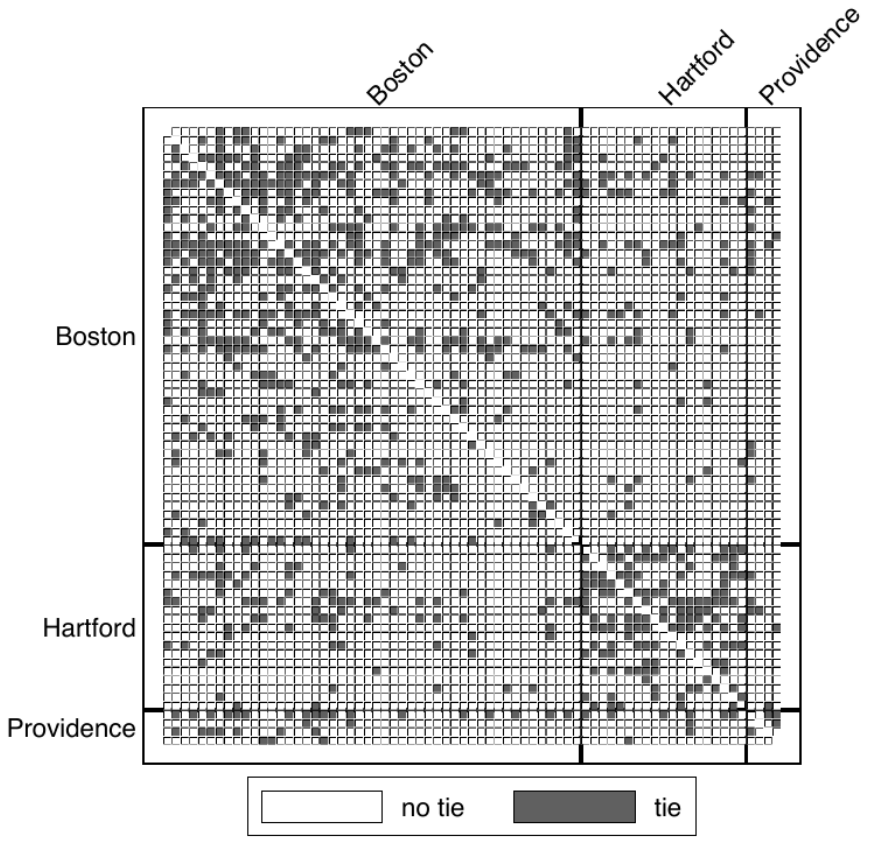
size(wealth): a measured variable, not an aesthetic choice. Same seed, same positions.

A matrix plot avoids the coordinate question altogether

```
. nwwebuse lazega, nwclear
. nwplotmatrix lazega_adv, group(office) ///
>         scheme(s1mono)
```

- A row is the lawyer who asks, a column the lawyer who is asked.
- Boston and Hartford show as dark diagonal blocks: advice stays within offices.
- The four Providence rows send most ties into the Boston columns.

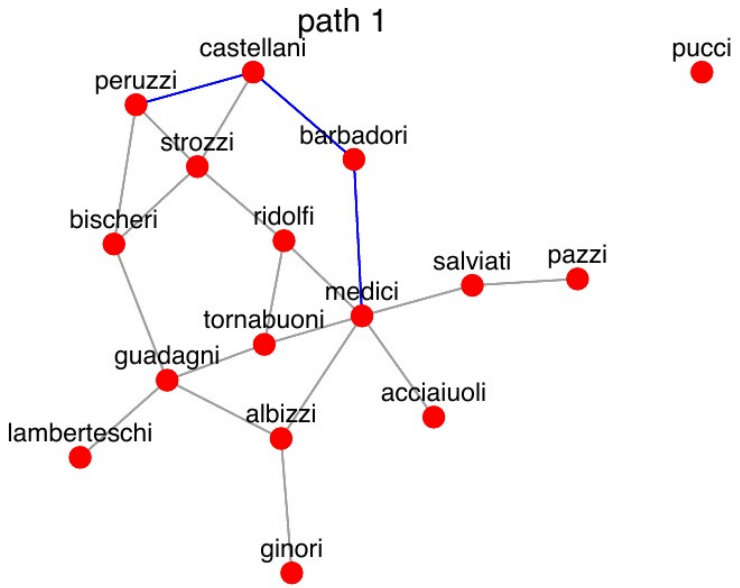
Weak spot: paths, and size. With 71 lawyers the matrix has 5,041 cells. Use it when the question is about blocks of an attribute, the interactive node-link page when it is about the shape of the network.

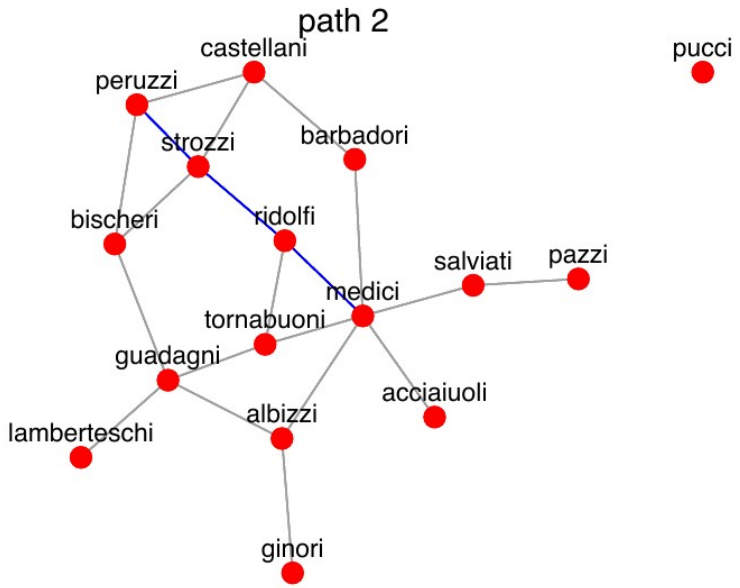


nwplotmatrix lazega_adv, group(office): nothing here was placed by an algorithm.

Highlight a result: show both shortest paths

```
. nwpath flomarriage, ego(medici) alter(peruzzi) generate(sp)
    Shortest path length : 3
  Path 1: medici <=> barbadori <=> castellani <=> peruzzi
  Path 2: medici <=> ridolfi <=> strozzi <=> peruzzi
. nwplot flomarriage, label(_nwnode) edgecolor(sp_1, legendoff)      // and edgecolor(sp_2, legendoff)
```

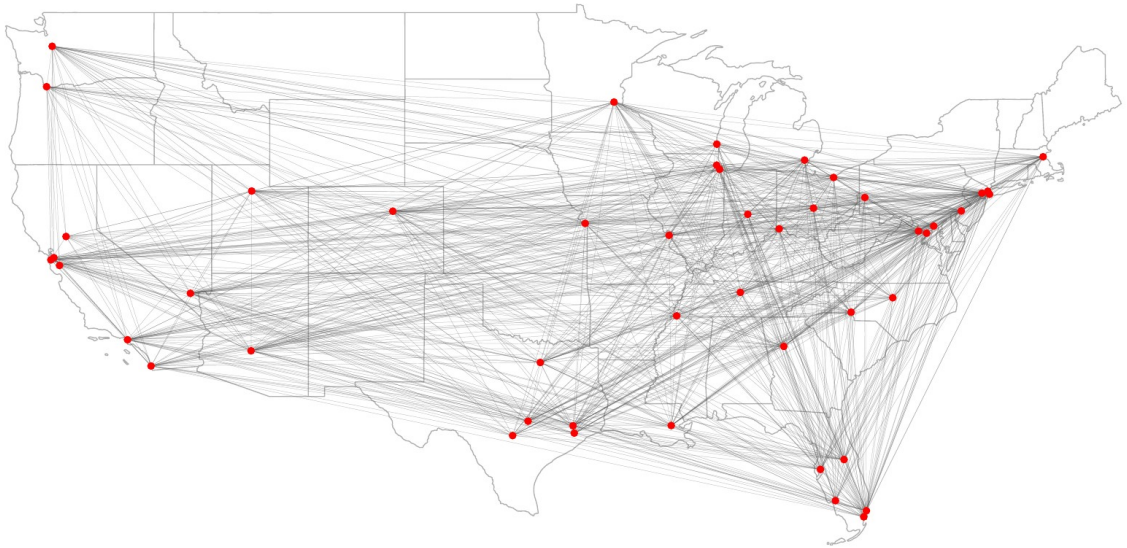




Where several paths tie, a single colored path looks like the only route. A highlight shows what the author chose to compute and hides everything else.

When real coordinates exist, the position is data

```
. nwwebuse usair, nwclear
. nwplot usair if Lat < 50 & Lon > -125, nodexy(Lon Lat) ///
>         basemap(usmap if !inlist(state, "AK", "HI")) arcstyle(straight) ///
>         nodefactor(0.3) edgefactor(0.1) lineopt(lcolor(gs6%15)) ///
>         arrowfactor(0) xsize(8) ysize(4.6)
```



- nodexy() uses longitude and latitude instead of a computed layout; basemap() draws the state outlines.
- Routes are drawn thin and transparent: where many run together, busy corridors show through.
- With real coordinates, moving a node on the interactive page destroys information: recolor and resize, but leave the nodes.

Explore interactively, then draw the figure: four steps

1 Explore

Write the page, open it in a browser; move, recolor, resize until you have seen what there is to see.

2 Save

Save what you keep as a do-file with the button Save do-file; it records the hand edits.

3 Decide

Each impression is a hypothesis: it dissolves when you move a node, or it names a measure to compute.

4 Draw

Run the do-file in Stata with the options print needs; say in the caption how the positions were obtained.

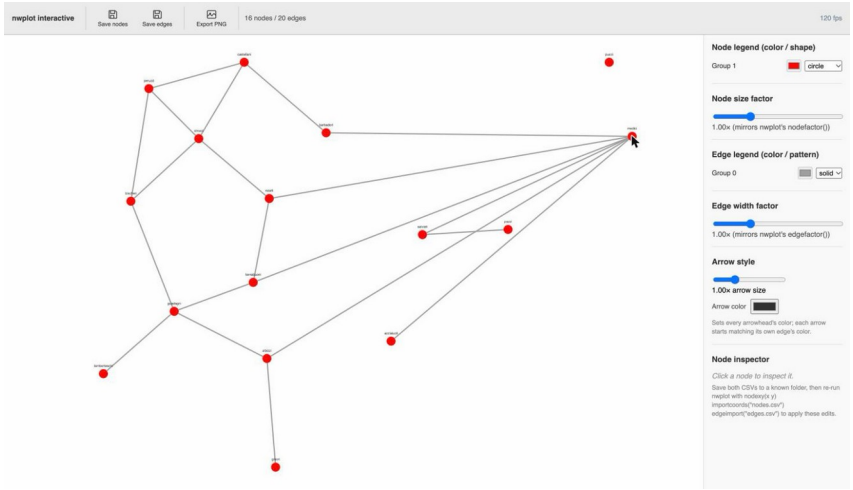
```
. set seed 42
. nwplot flomarriage, label(_nwnode) generate(x y) ///
>     interactive noopen fname(florentine_explore.html)
. run nwedit_plot.do          // after Save do-file on the page
```

The page computes nothing: no measure changes when you move a node. What you do on it is the analyst's choice, not data.

Interactive pages: open them and try

All pages are under nwcommands.org/book/interactive/

Page	What to try
`florentine/`	Drag Medici to the edge: its six ties stretch across the page; betweenness stays 47.5
`lazega/`	The chapter's workflow: color is office, shape practice, size indegree; recolor a group, move the sliders, Save nodes and Save edges
`glasgow-smoking/`	Glasgow friendship at wave 1, colored by smoking: drag the most conspicuous girl and bring the layout back with nodexy()
`chesapeake/`	Food web: why the edge-width slider cannot replace a log transformation



The page after one drag: Medici moved to the upper right.

Each page is a single file with its plotting library inside it: it runs in the browser, does not contact a server and does not need Stata.

Hands-on (20 min)

Ch. 5, Ex. 1

After nwwebuse florentine, nwclear, draw the marriage network flomarriage with label(_nwnode) in three layouts: the default kk, layout(circle) and layout(frucht), each after set seed 42. Name one visual feature that differs across the three pictures although the network is identical, and explain in one sentence why it is not evidence of anything structural.

Hint Look at where the Medici family sits relative to the edge of the picture, and where the isolated Pucci family is placed. Then check in the layouts table what each layout places nodes by.

Ch. 5, Ex. 3

Write the interactive page of the Florentine marriage network with fname(flo_drag.html) and generate(x y), open it in a browser and drag the Medici family to three different places. Say whether each changed: the number of Medici's ties, the lengths of its lines, its betweenness of 47.5, and the answer a reader would give to 'is Medici central?'. What does this show about reading layouts?

Hint The page moves positions only and computes nothing. Ask which of the four items could be computed from the ties without ever drawing the network.

What to take away

Ties are data; in most layouts positions are not

A claim about structure needs a measure defined on the ties. Medici's betweenness is 47.5 whether it sits in the middle or on the rim.

Explore interactively, then draw the figure for reporting

Keep the do-file, and say in the caption how the positions were obtained. Where positions are real coordinates, leave the nodes where they are.

Color, size and symbol carry information only when mapped to measured variables

A rank is not a size. Where positions are not coordinates, a matrix plot avoids the coordinate question altogether.

Day 1 in one line of work: from a question to a picture

Question

A relation on pairs

Chapters 1, 3

Construct

Boundary, tie, instrument, rule for one-sided reports

Chapter 3

Declare

nwset, nwfromedge, nwimport

Chapter 4

Check

nwsummarize: Directed, Nodes, density

Chapters 2, 4

Draw

nwplot, nwplotmatrix; positions are not data

Chapter 5

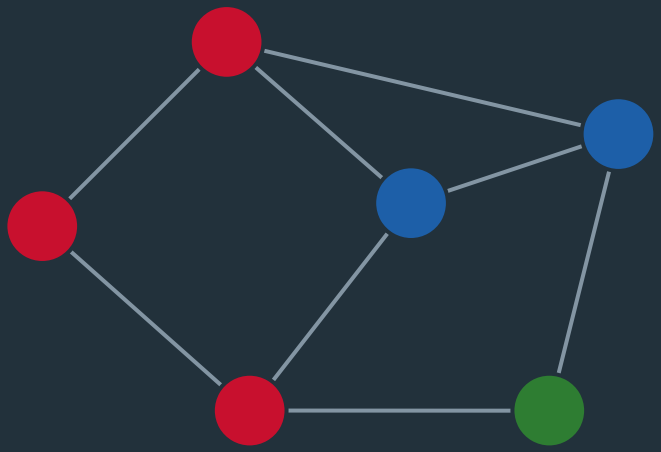
Tomorrow: Day 2

Chapters 6 to 8: describing a network (density, dyads, triads, mixing), paths and reachability, centrality and brokerage.

For tomorrow

Bring the filled-in checklist for your own data (Ch. 3, Ex. 10), and the three ideas of the day.

To read after the session



Read: Chapters 1 to 5; Chapter 6 for tomorrow

Run the do-files of Chapters 2 to 5 (nwcommands.org/book/dofiles) and the interactive pages (nwcommands.org/book/interactive/)

Exercises not done today: Ch. 1, Ex. 6; Ch. 2, Ex. 1 and 2; Ch. 3, Ex. 10; Ch. 4, Ex. 6; Ch. 5, Ex. 4 and 5

Do-files and teaching material: nwcommands.org/book

Instructors are free to use, adapt and share these slides.