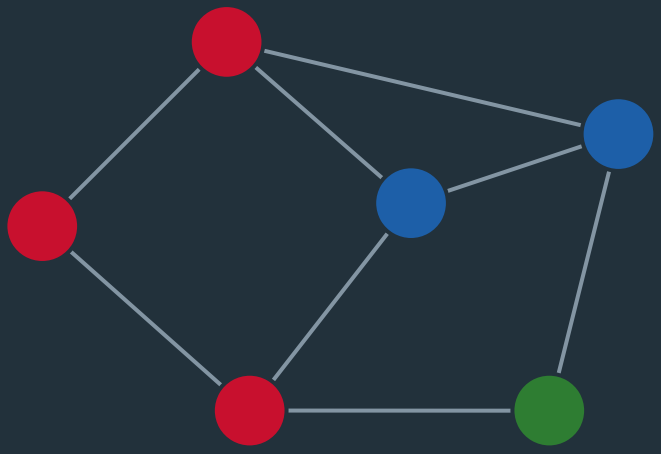
From a question to network data in Stata



Course A · Week 2 · Chapters 3 and 4

One lecture and one lab, about 90 minutes each. Network Analysis with Stata (Stata Press)

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

Week 2: a lecture, a short break and a lab

0:00	**1 A network is constructed, not found**	Lecture · Ch. 3	15 min
0:15	**2 Whose report counts**	Lecture · Ch. 3	15 min
0:30	**3 Declared once, found by name**	Lecture · Ch. 4	15 min
0:45	Break		10 min
0:55	**4 Lab: from reports to ties**	Lab · Ch. 3, 4	20 min
1:15	**5 Lab: declare, import, derive**	Lab · Ch. 4	30 min

One question, one firm, three ideas for the week

The running question

How is informal advice-seeking organized within the firm, and does it follow the firm's formal hierarchy or cut across it? Seventy-one lawyers, three offices, three relations (Lazega).

Where Week 1 stopped

892 advice arcs, 700 edges after symmetrizing. But who decided that these 71 people, this relation and this rule are the network?

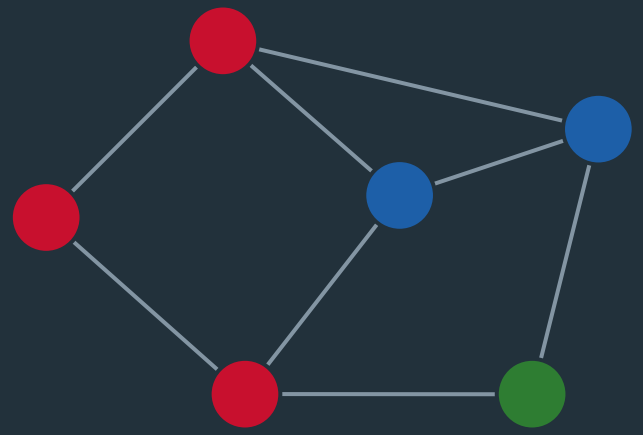
Three ideas

- A network is constructed, not found: boundary, tie, instrument, reports, missing data and time each shape what it can answer.
- The same survey answers give 501 or 74 friendship ties, depending on a rule that belongs in every report.
- In Stata a network is declared once and then found by name; a wrong declaration produces no error, only a wrong network.

1

A network is constructed, not found

15 min · Chapter 3 · six decisions



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

Choice	Researcher A	Researcher B
How many friends	Each pupil names up to three	A class list, a yes or no for every name
A one-sided nomination	Does not count: only pairs who named each other	Counts: every pair named by at least one
The resulting network	Sparser, more conservative	Denser

Nobody made an error

Neither network is false. The instrument shapes the result: how many names, whether a one-sided nomination counts, who is inside the school, how often it is asked.

An operationalization

A network is to a social fact what a survey item is to an attitude. The classic mistake is to treat the edge list in the file as the thing itself.

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: an affective bond or recent interaction? Directed? Valued?

Instrument

Name generator, with or without a ceiling on the 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?

For the firm: advice is directed and the question is about who asks whom, so the directed network is kept and no reconciliation rule is needed. A rule matters once a relation is reduced to an undirected one.

The boundary decides which measures mean anything

Realist boundary

Follows the actors' own sense of who belongs: a club, a gang, a group that sees itself as one.

Nominalist boundary

Imposed by the researcher: a membership list, everyone named as a collaborator on a set of papers, everyone within two steps of a seed set.

- Distance and betweenness are computed inside the boundary and say nothing about paths that leave it.
- Named actors who are not surveyed can stay as nodes with no further data (the boundary is the sampled population), or a snowball design can follow them (the boundary keeps expanding until someone stops it).
- What matters is that the decision is made and stated, not left to whatever the instrument happened to do.

Reading the boundary off the data: 71 lawyers, three offices

```
. 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

48 lawyers in Boston, 19 in Hartford, 4 in Providence; 35 associates and 36 partners. Most Hartford lawyers are partners.

What it cannot show

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

The tie and the instrument shape the degrees

	Name generator	Roster
Asks	'Who do you turn to for advice?' The respondent supplies names, up to a ceiling or without one	A list of all possible alters; a relation to each
Strength	Scales to large, unbounded populations	No recall problem: nobody forgets a colleague on the list
Weakness	Recall suffers; a ceiling cuts the top of the degree distribution	Burden grows with the population; late names tend to get a 'no'

Which relation is asked matters as much: 'Do you consider this person a friend' and 'Have you spoken to this person in the last week' give very different networks among the same people. Randomize the order of a roster across respondents.

6

friends per pupil allowed in the Glasgow study

12

classmates allowed in the klas12b study

A pupil with exactly six nominations may 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 and, at most, the ties among them
Reaches	Bounded groups: a firm, a class, a gang	Large or loosely bounded populations: a national sample
Cannot give	Large populations at low cost	Closeness, betweenness, communities (Chapters 8, 9)

Missing data is not a missing cell

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

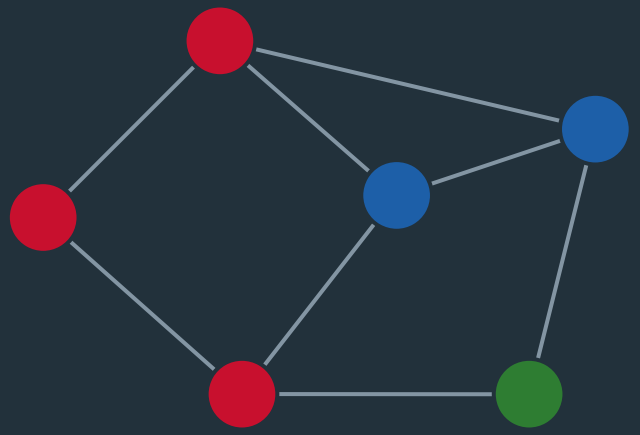
A response rate is not a footnote

People who skip a survey are often the isolated or the unpopular, so the bias is systematic. Report the rate and how missing ties were handled.

2

Whose report counts

15 min · Chapter 3 · AND, OR or directed



Three conventions turn the same reports into three networks

Pair of respondents	AND rule	OR rule	Kept directed
Both name each other	1 tie	1 tie	2 arcs
Only one names the other	no tie	1 tie	1 arc
Neither names the other	no tie	no tie	no arc

AND: mode(min)

Keeps a tie only if both report it. Sparser and more conservative; discards every one-sided report.

OR: mode(max)

Keeps a tie if either reports it. Denser; counts a claim the other person might not recognize.

Directed

Keeps each report as it was made. The asymmetry is information: for advice, the question is who asks whom.

Friendship among 71 lawyers: 501 edges 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): 74 + 427 pairs

74

AND rule, mode(min): the mutual pairs

575

arcs, kept directed

It says

The same survey answers give about 6.8 times as many ties under the OR rule as under the AND rule. In 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.

Three densities, three denominators

Network	Ties	Possible	Density
Directed friendship	575 arcs	4,970 ordered pairs	.116
OR rule	501 edges	2,485 unordered pairs	.202
AND rule	74 edges	2,485 unordered pairs	.0298

The denominators differ: ordered pairs for the directed network (71 · 70 = 4,970), unordered pairs for the two undirected ones (71 · 70 / 2 = 2,485). A difference between these densities is not a difference in how connected the firm is.

A write-up that reports 'the friendship network of the firm' without naming its rule leaves the reader unable to tell which of the three it describes.

It says
Under the OR rule 501 of the 2,485 pairs are tied; under the AND rule only 74.

It does not say
That the firm is more cohesive under one rule than under another. .116 divides by ordered pairs and .202 by unordered pairs, so the two are not on one scale.

The reporting checklist: six questions

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

The checklist applied to the lawyers' data

	Question	The lawyers' data
1	Boundary	All members of one firm, three offices: a positional boundary. Outside advisers 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 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.

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 (Chapters 15 to 18), then permutation tests and QAP (Chapters 19, 20).

Model

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

- The method follows from the question, the data structure and the goal, not from the availability of a command.
- A centrality score is evidence of a position, not of influence; a significant coefficient is evidence against a stated null, not of a cause.

What to take away from Chapter 3

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 edges under the OR rule and 74 under the AND rule; the rule belongs in every description of the data.

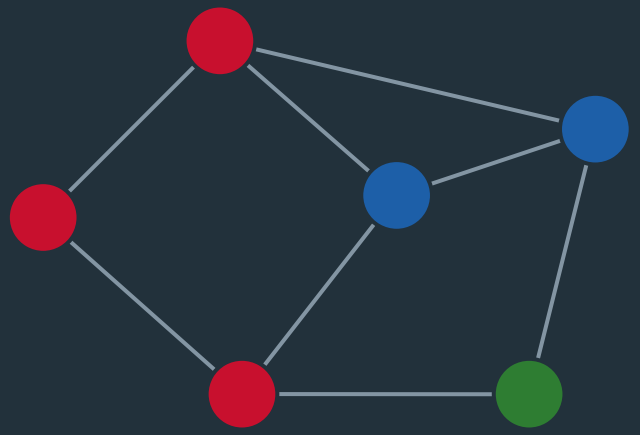
Choosing a method is the second half of the same design

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

3

Declared once, found by name

15 min · Chapter 4 · the idea; the commands follow in the lab



Networks sit beside the data and are found by name

```
. nwwebuse lazega, nwclear
. nwset
(3 networks)
--------------------
      lazega_adv
      lazega_cow
      lazega_fr
```

Relational data break the assumption that a row does not depend on any other row: the unit of observation is the pair of actors.

The data: ordinary variables

71 rows, one per lawyer, 13 variables: the node label, an inclusion flag and eight attributes, with labeled versions of three of them.

The networks: objects

Three relations among the same 71 lawyers, stored beside the data and found by name. A command takes a name, a wildcard or a list.

Three shapes of data, three ways in

The data are	Command	Watch for
A list of ties, one row per tie	nwset ego alter, edgelist	directed or undirected must be stated; a row with an empty alter keeps an isolate
A matrix of ties	nwset, mat()	isolates are kept; labs() names the nodes
A file from Pajek, UCINET, GML or a text edge list	nwimport	type() and name(); check the inferred Directed: line
One row per respondent, one column per nomination	reshape long, then nwfromedge	choose directed or undirected deliberately
A network to keep between sessions	nwsave, nwuse	the network analogue of save and use

The declaration of directedness, once typed, is stored: every later command reads it silently.

An edge list forgets people who have no tie

```
. list, clean
          id   nomina~n   friend
  1.    anna          1   sophie
  2.    anna          2     gwen
  3.    fion          1     gwen
  4.    fion          2   sophie
  5.    gwen          1     anna
  6.    gwen          2      leh
  7.    gwen          3     fion
  8.     leh          1
  9.  sophie          1     anna
```

The survey of the lab after reshape long and the chapter's drop: row 8 keeps Leh, who names nobody, with an empty receiver.

The edge list: isolates vanish

A person nobody names who names nobody appears in no row. n shrinks, density and every average inflate, and the vanished are the isolated, the ones a study of exclusion cares about. Remedy: a row with a sender and an empty receiver.

The matrix: every actor owns a row

No isolate can be lost. The limit is size: a cell for every pair, wasteful for a large sparse network.

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)
   status |    adv_in   adv_out
----------+--------------------
associate |  9.428571  7.885714
  partner |  15.61111  17.11111
----------+--------------------
    Total |  12.56338  12.56338
```

15.6

mean advice ties received by partners, against 9.4 for associates

17.1

mean advice ties sent by partners, against 7.9

It says

Partners are asked more and ask more than associates. Both degrees are ordinary variables now, so tabstat, graph, regress and margins work on them.

It does not say

Why. Position, selection and a shared cause are all still open: the rival accounts of Chapter 1. Nor that the two means are equal in total: mean in and mean out are identical in every directed network.

Vary the decision: new networks from old ones

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

No argument settles every decision; the remedy is to vary it and see whether the answer moves.

The originals stay

Each version gets its own name. The same analysis runs on several versions, and the report says how much the answer depends on the choice.

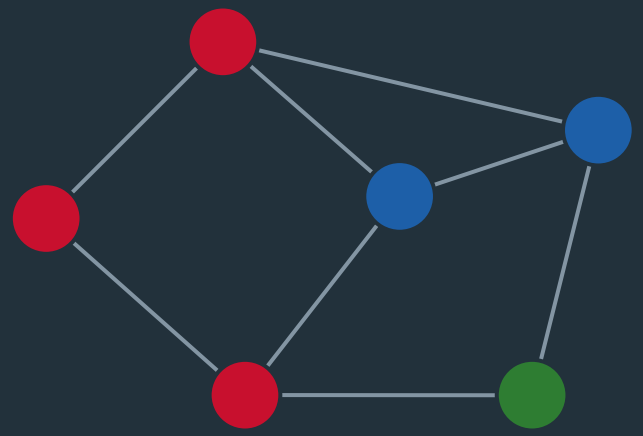
Chapter 3 meets Chapter 4

The boundary becomes nwsubset, the reconciliation rule becomes mode(max) or mode(min).

Break

10 minutes

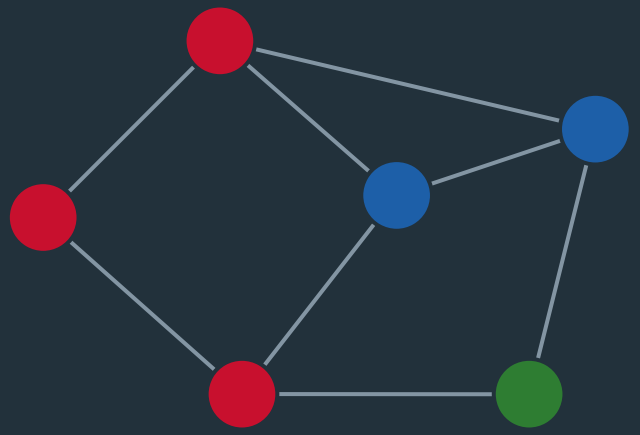
Back for the lab



4

Lab: from reports to ties

20 min · Chapters 3 and 4 · type along, then the exercises



A survey export: reshape long gives one row per nomination

```
. nwclear
. clear
. 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
```

5 → 15 → 9

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

- **nomination > 1**

  keeps the first slot of anyone who named nobody, here Leh, as a row with an empty receiver.

From nominations to ties: undirected applies the union rule

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

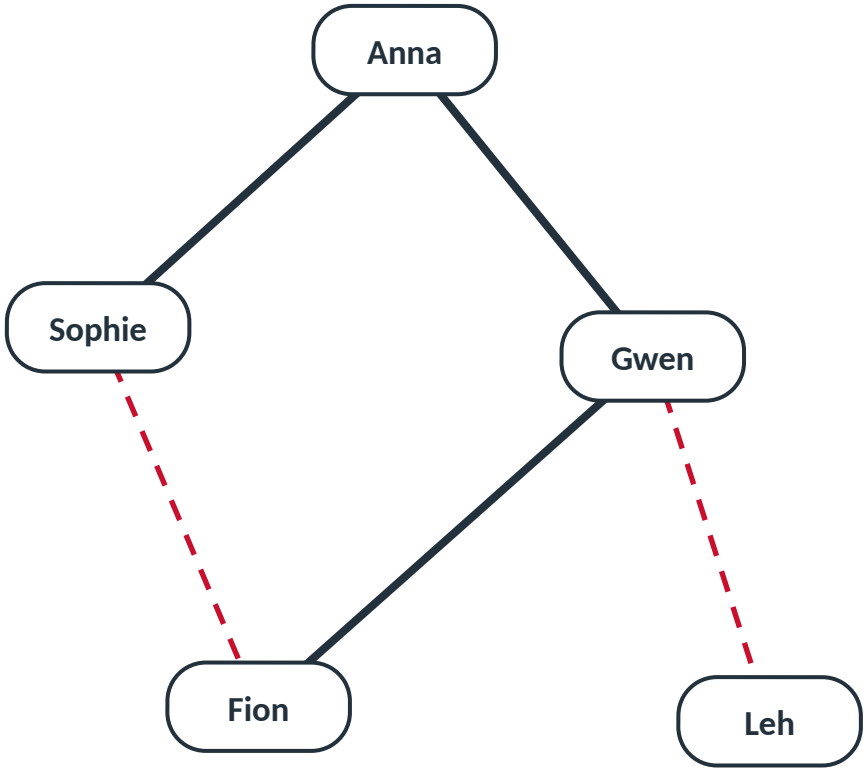
8 → 5

nominations become undirected ties

3

ties under the intersection rule (the solid ones)

Density .5 is 5 of 10 possible ties; under the intersection rule 3 of 10 = .3. The option undirected has silently chosen the union rule.



Solid: named from both ends. Dashed: named from one end only.

The same two rules on the lawyers' friendship network

```
. nwwebuse lazega, nwclear
. nwsym lazega_fr, generate(fr_or) mode(max)
. nwsym lazega_fr, generate(fr_and) mode(min)
. nwsummarize fr_or fr_and
    Network name:  fr_or
    Edges: 501
    Density:  .202
    Mean degree:  14.11
    Network name:  fr_and
    Edges: 74
    Density:  .0298
    Mean degree:  2.08
```

It says

The two networks reproduce the counts of the lecture: 501 and 74 edges, a mean degree of 14.1 against 2.1, depending on a single option.

It does not say

A choice between them. Which rule is right depends on what a tie must stand for; the lecture's question, who counts as a friend, is a question about the relation, not about Stata.

Hands-on (20 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

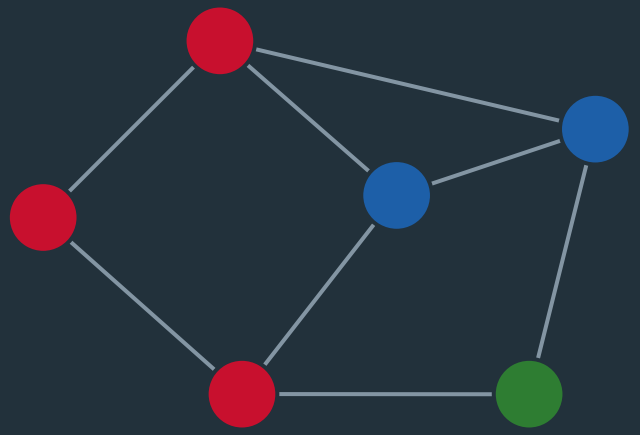
The chapter applied the OR and AND rules to friendship (501 against 74 edges, a ratio of about 6.8). Do the same for advice: after nwwebuse lazega, nwclear, run nwdyads lazega_adv, predict the number of edges under each rule from the dyad census, then check with nwsym lazega_adv, generate(adv_or) mode(max), nwsym lazega_adv, generate(adv_and) mode(min) and nwsummarize adv_or adv_and. Report both edge counts, both densities and the 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.

5

Lab: declare, import, derive

30 min · Chapter 4 · type along, then the exercises



Declare from a matrix: nwset, mat()

```
. 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)
```

Row i, column j holds the arc from i to j. Rows are separated by a backslash; labs() names the nodes in the order of the rows; directed is typed once and stored.

The matrix holds the same four arcs as the edge list of Chapter 2 (row Gwen has two entries, row Leh none): a network does not change when it is written down in another format.

Start from an empty session

nwclear removes the networks of earlier work. Stata's clear removes the data, not the declared networks.

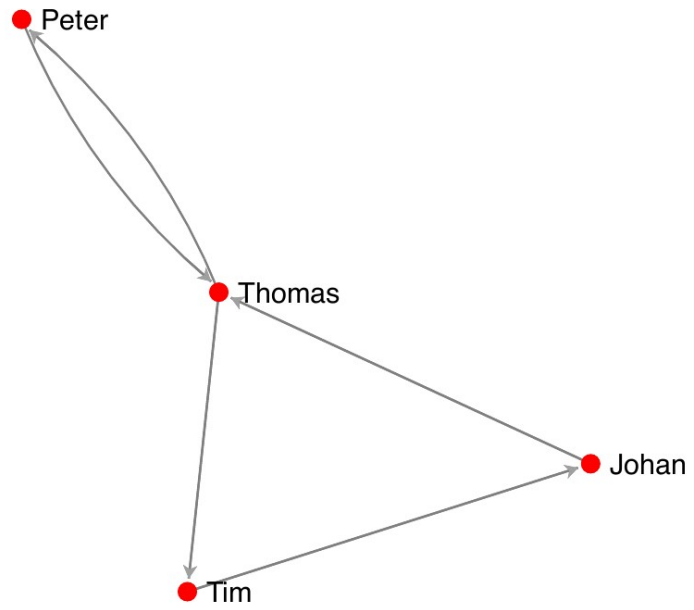
Outdegrees 1, 1, 2, 0

Anna and Sophie send one arc, Gwen two, Leh none: the same degree tables as in Chapter 2.

Import: nwimport reads a file from a folder or an address

```
. 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: the Thomas to Peter and Peter to Thomas rows are recorded separately. Check the Directed: and Valued: lines against what you know about the relation.



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

Several networks, one data set: 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
```

* 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.

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

892 advice arcs (density .179), 1,104 work arcs (.222), 575 friendship arcs (.116): three relations among the same 71 lawyers.

Between sessions: nwsave and nwuse

```
. 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. The .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. Stata/BE allows 2,048 variables, so a very large network cannot be loaded this way.

Derive networks from networks: subset and symmetrize

```
. nwwebuse lazega, nwclear
. nwsubset lazega_adv if office == "Boston", name(adv_boston)
. nwsummarize adv_boston
   Nodes: 48
   Arcs: 539
   Density:  .239
. nwsymmetrize lazega_fr, mode(max) generate(fr_or)
. nwsymmetrize lazega_fr, mode(min) generate(fr_and)
. nwset
(6 networks)
```

The pairs across offices, which are tied less often, are no longer counted. The two symmetrized networks are the 501 and 74 edges of the lecture (density .202 and .0298): one option decides.

539

of the 892 advice arcs stay among the 48 Boston lawyers: density .239 against .179

The originals stay

nwset now lists six networks: the three lawyers' networks, adv_boston, fr_or and fr_and. Every derived network has its own name.

Hands-on (25 min)

Ch. 4, Ex. 1

Declare a new five-person network from an adjacency matrix typed with nwset, mat(), as the toy network was declared: 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 by hand the number of nodes, the number of arcs and the density, 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 and labs() gives the names in the order of the rows. 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. 6

Return to the survey example (with Gwen naming Leh, as in the chapter). Build the network as directed, nwfromedge id friend, name(fdir) directed, and report its number of arcs and its density. Then derive the two undirected versions with nwsymmetrize: mode(max) for the union rule and mode(min) for the intersection rule. Report the edges and the density of each, and say which version the chapter's undirected option produced.

Hint Start from the long data again, up to and including the drop: nwfromedge replaces the data in memory by the network, so the edge list is gone after the first call. Then nwsymmetrize fdir, mode(min) generate(f_and), the same with mode(max) and generate(f_or), and nwsummarize fdir f_and f_or. A directed network of five people has 5 · 4 = 20 possible arcs, an undirected one 10 possible edges.

What goes wrong, and how the output shows it

A nomination treated as a tie

undirected is one word and the output is tidy, but it applies the union rule silently: five ties from eight nominations.

Isolates lost

A list of ties holds only people who appear in a tie. Give each other person a row with an empty receiver, or declare the network from a matrix.

An inferred property trusted

A correct guess by nwimport does not make the next file safe: check the Directed: and Valued: lines of nwsummarize against the design.

clear does not clean the registry

It removes the data in memory, not the declared networks. Use nwclear before declaring a network under a name that may already exist.

What to take away this week

A network is constructed, and the construction is part of the result

Boundary, tie, instrument, whole or ego design, missing data and time shape what a network can answer. The rule that turns reports into ties changes the count several times over (501 against 74), and belongs in every description.

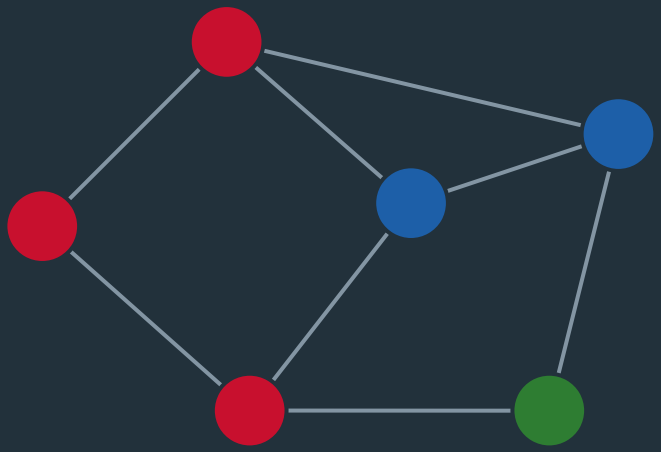
In Stata a network is declared once and then found by name

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

Vary a decision, do not argue it

nwsubset, nwsymmetrize and their relatives derive new networks under new names from the declared one, so an analysis can be run on several versions and the answer's dependence on the choice reported.

Before next week



Read Chapter 5, Seeing networks: layouts, node attributes in a picture, the matrix plot.

Left over from this week: Ch. 3, Ex. 1 (two studies of one office) and Ex. 8 (OR or AND for two questions); Ch. 4, Ex. 4 (wildcards, nwsave, nwuse) and Ex. 5 (work and friendship overlap).

Bring a relation from your own field and answer the six questions of the reporting checklist for it (Ch. 3, Ex. 10).

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

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