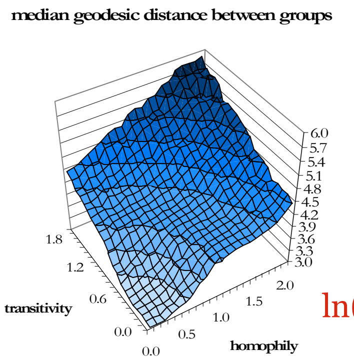
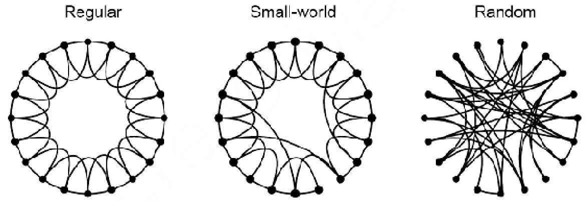
university of groningen / behavioural and social sciences / sociology

Statistical Analysis of Complete Social Networks

Introduction to statistical inference for complete networks: classical approaches

Christian Steglich c.e.g.steglich@rug.nl





$$\ln\left(\frac{\Pr(x^c \rightarrow_i x^b)}{\Pr(x^c \rightarrow_i x^a)}\right) = \sum_{k=1}^{K} \beta_k \left(s_{ik}(x^b) - s_{ik}(x^a)\right)$$

Interdependence of observations

We will see on the next slides examples for...

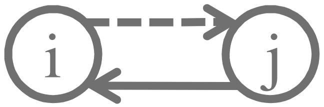
- ...interdependence of tie variables;
- ...interdependence of tie and actor variables.

But, what is an "observation" in a complete network study?

Ultimately, the whole network should be treated as *one* (admittedly very complex) *observation*, with intrinsic dependencies among its constituent elements.

Interdependence of tie variables (1)

Tie variables $\{x_{ij} \mid i, j \in \{1,\dots,n\}, i \neq j\}$ in a complete network data set tend to exhibit certain dependencies:

› *reciprocity* 

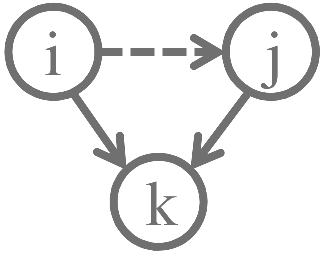
In egalitarian networks (friendship, trust, communication), $\mathbf{x_{ij}=1}$ is *more* probable when $\mathbf{x_{ji}=1}$ than when $\mathbf{x_{ji}=0}$.

In hierarchical networks (power attribution, supply chains), $\mathbf{x_{ij}=1}$ is *less* probable when $\mathbf{x_{ji}=1}$ than when $\mathbf{x_{ji}=0}$.

$$\text{reciprocity index} = \frac{2M}{2M+A} \neq \frac{2M+A}{n(n-1)} = \text{network density}$$

Interdependence of tie variables (2)

› *structural equivalence*



Probability of $\mathbf{x_{ij}=1}$ can depend on identity $\mathbf{x_{ik}=x_{jk}}$.

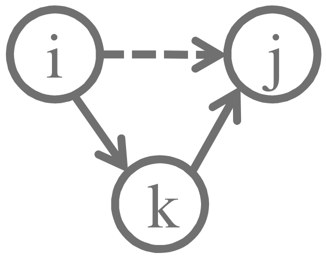
If two actors **i,j** are identically tied to all third parties **k**, this identifies them as "structurally equivalent".

In affective networks (friendship), structural equivalence tends to *facilitate* a direct relation. In instrumental networks (trade) it tends to *inhibit* it (structural competition).

Measures make use of the tie correlation $\sum_{ijk} x_{ij} \operatorname{cor}_k(x_{ik}, x_{jk})$ or the raw equivalence count $\sum_{ijk} x_{ij} \left(x_{ik} x_{jk} + (1 - x_{ik})(1 - x_{jk}) \right)$

Interdependence of tie variables (3)

› *transitivity*



Probability of $\mathbf{x_{ij}=1}$ can depend on co-occurrence of $\mathbf{x_{ik}=1}$ and $\mathbf{x_{kj}=1}$.

'Friends of my friends are my friends'.

Transitivity measures hierarchy-compatible group formation in a (social) network.

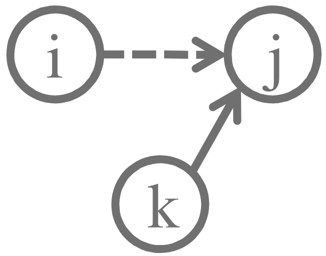
It overlaps with structural equivalence!

Possible measure: transitivity index

$$\left(\sum_{ijk} x_{ij} x_{ik} x_{jk}\right) \Big/ \left(\sum_{ij} x_{ij} x_{ik}\right)$$

Interdependence of tie variables (4)

› *degree differentials*



Probability of $\mathbf{x_{ij}=1}$ can depend on $\sum_{k \neq i} \mathbf{x_{kj}}$.

'Matthew effect': the rich get richer.

Many networks have 'hubs'.

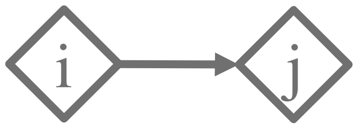
Again, there is overlap with other constructs – here transitivity.

› *etc. – many such 'endogenous dependencies'*

Interdependence of tie and actor variables

Tie variables $\{x_{ij} \mid i, j \in \{1,\ldots,n\}, i \neq j\}$ and actor variables $\{z_i \mid i \in \{1,\ldots,n\}\}$ tend to exhibit certain dependencies:

› *homophily*



‘Birds of a feather flock together’: In many networks, $\mathbf{x_{ij}=1}$ is more probable when $|\mathbf{z_i - z_j}|$ is small than when $|\mathbf{z_i - z_j}|$ is large.

Homophily *induces reciprocity* (because $|\mathbf{z_i - z_j}| = |\mathbf{z_j - z_i}|$).

Classical measures are *network autocorrelation indices* like

Moran’s $I = \dfrac{n \sum_{ij} x_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\left(\sum_{ij} x_{ij}\right)\left(\sum_i (z_i - \bar{z})^2\right)}$ or Geary’s $c = \dfrac{(n-1)\sum_{ij} x_{ij}(z_i - z_j)^2}{2\left(\sum_{ij} x_{ij}\right)\left(\sum_i (z_i - \bar{z})^2\right)}$

Lessons learnt

- There are multiple types of dependencies (dyadic and triadic in nature, potentially of higher order) among network ties.
- These can (and do) co-occur in empirical data.
- They often constitute qualitatively different "social mechanisms" / explanations of theoretical interest (e.g., reciprocity norms vs. homophily).
- Aim of a statistical approach should be to *express* them, maybe *separate* and *identify* them, certainly *control* for their occurrence in network data.

Hypothesis tests for network data

‘Classical SNA’ *is mainly about descriptive network statistics*

- proximity, similarity, centrality, brokerage,...
- positional measures, equivalence,...

Hypothesis testing *requires an inferential-statistical approach*

- Crucial are meaningful *distributions* of test statistics, on which *p-values* for hypothesis tests can be based.
- It is not trivial to construct such *“meaningful distributions”* for complete network data!

Examples of research questions:

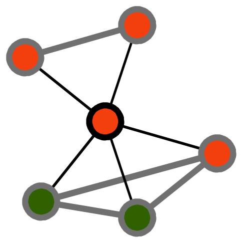
- *In a dynamic network, do central actors emerge by pure network-inherent, structural processes – or do personal characteristics 'predestine' some actors towards centrality?*
- *Do close friends have more influence than other friends, on individuals' alcohol consumption, political opinion, music listening habits, obesity, etc.?*
- *For interpersonal conflict at the workplace, do differences in work attitude matter more than spatial proximity?*

Such questions can easily be reformulated as hypotheses to be tested on network data.

Complete vs. ego-centered data, revisited

Ego-centered data:

- A random sample of actors is drawn,
- each of the actors' network neighbourhood is measured.

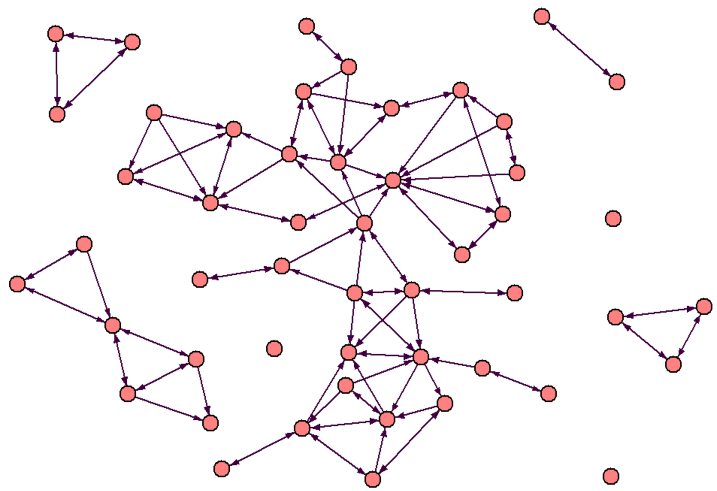


many small networks

Complete data:

- A group of actors is decided on,
- all network ties existing in this group are measured.

one bigger network



Statistical tests for ego-centered network data

- Data on the actor level have probability distribution of random samples:
  - 'classical' statistical techniques (regression, ANOVA,...) are possible for such data
  - typical research question: "Is local clustering of the ego-network related to ego's performance?"
- Data on the dyad level have multilevel structure:
  - nominated alters are 'nested' in the nominating egos
  - multilevel analytical techniques are possible
  - typical research questions: "Does the intensity of the relation between ego and alter depend on alter's performance? Does it depend on the number of network partners ego has?"

Statistical tests for complete network data

For many research questions, studying complete network data is expedient:

- Studying individual properties of actors and dyads:
  - Some individual-level network variables depend on more than immediate neighbourhood:
    - social capital, centrality, 'role positions',...
- Studying the network on its own behalf:
  - Macro structure can reveal properties of the social system that are barely visible at the actor level:
    - clustering, social balance, core-periphery structures,
    - small world phenomenon, segregation,...
- Studying selection processes:
  - You need to know about pool of eligible partners (also non-chosen ones) to find out what drives partner selection.

Complete network data are special:

- Sampling dependence of actors:
  - A complete network study always relies on measures of all actors in a given social context, **not** on random samples.
- Structural interdependence of dyads:
  - Two relationships involving the same actor are likely affecting each other.
- Higher-order dependence (think of examples given above):
  - Absence of a relational tie between two actors may increase likelihood for third actors to function as bridge between them.

Whenever complete network research is meaningful,
*there is **no** independence of observations.*

Necessary for hypothesis testing are …

- a test statistic operationalising the hypothesis,
- the distribution of the test statistic according to a null hypothesis / null model.

Then, a *p*-value can be calculated indicating likelihood of the observed value of the test statistic (or a 'more extreme' value) under the null model.

Typically the null distribution is based on the sampling process (sampling distribution).

For complete networks, this is not possible!

Reminder / Excursion: Inference based on sampling distributions

› A good sampling process induces a *reference distribution* for sample statistics.

› Simple Random Sample:
  - Each sample of same size has same probability;
  - Assumption about population plus sampling process implies distribution of sample statistic;
  - Test statistics can exploit the known properties of this sampling distribution.

Standard example: The sampling distribution of the sample mean

› Suppose a numerical variable X has a population mean μ and a population standard deviation σ.

› Suppose we study the space of all simple random samples of size n drawn from this population.

› <u>*Central limit theorem:*</u> The sample mean $\bar{X}$ in this sample space approximately (for large n) follows a *normal distribution* with mean μ and st.dev. $\sigma / \sqrt{n}$.

⇒ The test statistic $t = \bar{X} / (s / \sqrt{n})$ approximately follows a *Student's t distribution* with $n-1$ degrees of freedom.

Alternatives to sampling distributions

"Non-parametric" alternatives

1. Distributions assuming *tie* or *dyad* independence, or other forms of *conditional* independence;
2. Distributions under *permutations* of the actor labels.

Hybrid models (secondary analyses using SNA descriptives)

3. Distributions of actor or dyad variables assuming *conditional independence, given the results of a primary, descriptive social network analysis*.

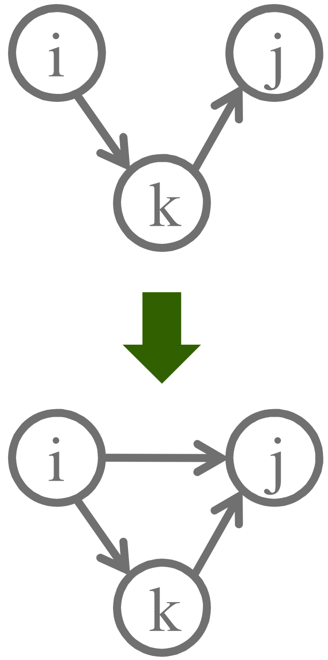
Model-based ("parametric") alternatives

4. Distributions derived from *explicit models* of the dependence between actors / dyads.

1. Distributions assuming tie independence

Example question:

Is there evidence for transitive closure in a given network?



…could be operationalised

1. by counting transitive triplets (configuration on the bottom), or
2. by calculating the fraction of transitive triplets among both configurations, or by still other quantities.

Expected values of these statistics under the assumption of tie independence ("null hypothesis"):

1. Expected count of transitive triplets is

$$E(T) = E\left(\sum_{ijk} x_{ij} x_{jk} x_{ik}\right)$$
$$= \sum_{ijk} \Pr(x_{ij} = 1 \wedge x_{jk} = 1 \wedge x_{ik} = 1)$$
$$= n(n-1)(n-2)p^3$$

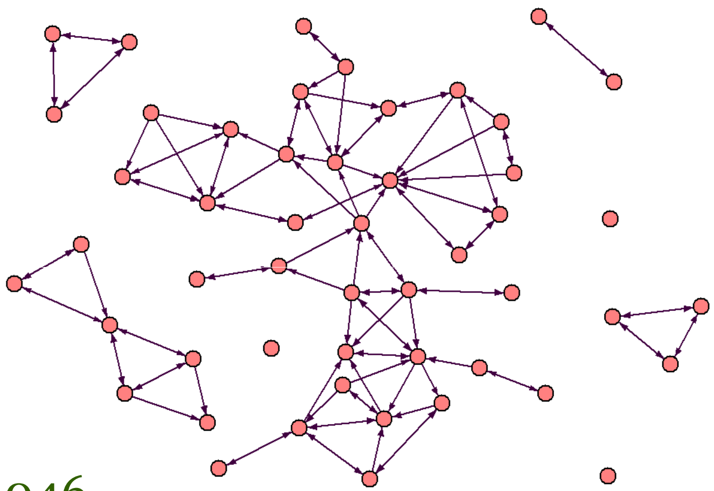
distribution is binomial B(n(n-1)(n-2),p^3).

2. Expected fraction of transitive triplets among both configurations is

$$E(f_T) = p$$

distribution is binomial B(n(n-1),p)/(n(n-1)).

For the complete network shown (1st observation of the s50 network), the observed values are these:



50 actors
113 network ties, density 0.046
86 transitive triplets
136 intransitive triplets
222 'configurations of interest'

1. Expected count of transitive triplets is ~11.25, p-value is far below 0.0001.

2. Expected fraction of transitive triplets is 0.046, observed fraction is 86/222=0.387, p-value again is far below 0.0001.

Results are typical: tie independence is a bad (a priori unrealistic) null model.

Other "conditional uniform" tests

A test assuming *dyad* independence (more realistic than tie independence) was proposed for the triad census of the network by Holland & Leinhardt (1978).

- This is the so-called "conditional uniform test, given the dyad census", or U|MAN test.
- Faust (2007, 2010) showed that conditioning on the dyad census, around 90% of the variance in triad distributions can be explained.
- …but then, is triad variance a meaningful yardstick?

Output U|MAN test (obtained with the `Pajek` software) :

Triadic Census.

Type	Number of triads (ni)	Expected (ei)	(ni-ei)/ei
1 - 003	16243	14764.27	0.10
2 - 012	1470	4283.34	-0.66
3 - 102	1724	103.56	15.65
4 - 021D	5	103.56	-0.95
5 - 021U	18	103.56	-0.83
6 - 021C	21	207.11	-0.90
7 - 111D	42	10.01	3.19
8 - 111U	30	10.01	2.00
9 - 030T	5	10.01	-0.50
10 - 030C	0	3.34	-1.00
11 - 201	15	0.24	60.96
12 - 120D	6	0.24	23.78
13 - 120U	5	0.24	19.65
14 - 120C	2	0.48	3.13
15 - 210	9	0.02	383.40
16 - 300	5	0.00	26498.63

Chi-Square: 164896.9327*** ← **independence hypothesis is rejected**

Warning:
7 cells (43.75%) have expected frequencies less than 5.

Alternative null distributions that have been studied in the same tradition control – instead of the dyad census – for the *degree distributions*: Snijders (1991), Karlberg (1999).

Problem with the whole approach:

When testing structural properties, the null hypothesis of conditional independence is pretty much always rejected.

So why continue to work with it at all?

...one reason may be to convince network-reluctant editors and reviewers of the necessity to do 'real' network modelling!

2. Distributions under permutations of the actors

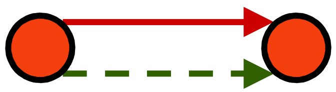
Basic idea (is somewhat similar to bootstrapping):

- Necessary: At least two variables that contribute to the test statistic.
- Re-use the (non-random) sample to generate a null distribution of the test statistic
- by permuting the actors and
- calculating one variable's contributions to the test statistic based on permuted actors' values, while
- calculating the other variable's contributions based on unpermuted values.

Advantage: Univariate distributions and network structure are invariant under permutations – i.e., controlled for!

Example question:

Is there evidence for an association between friendship and communication?



...could be operationalised

1. by counting the joint occurrence of ones in the two adjacency matrices, or
2. by calculating the Pearson correlation coefficient, taking the n(n-1) cells in the adjacency matrices as units of analysis [but note: for binary data, this is not the most suitable measure of association – better would be a 'phi coefficient' controlling for marginal frequencies].

Output (from UCINET):

```
QAP CORRELATION
--------------------------------------------------------------------------------

Data Matrices:                C:\1 Wien 2007\data sets used\MBA\Communication1Ril
                              C:\1 Wien 2007\data sets used\MBA\Friendship1Ril
# of Permutations:            5000
Random seed:                  24322

QAP results for C:\1 Wien 2007\data sets used\MBA\Friendship1Ril *
C:\1 Wien 2007\data sets used\MBA\Communication1Ril (5000 permutations)

          Obs Value Significa  Average   Std Dev   Minimum   Maximum   Prop >= O Prop <= O
          --------- ---------  --------- --------- --------- --------- --------- ---------
Pearson
Correlation 0.485      0.000     -0.000     0.021    -0.069     0.078     0.000     1.000

QAP Statistics

QAP Correlations         Commu Frien
                         ----- -----
    Communication1Ril    1.000 0.485
       Friendship1Ril    0.485 1.000

QAP P-Values             Commu Frien
                         ----- -----
    Communication1Ril    0.000 0.000
       Friendship1Ril    0.000 0.000
```

The significance level is the probability to exceed the observed value of 0.485 in the permutation-based calculations.

The null hypothesis is rejected.

Two other important permutation-based tests:

Moran's **I** and Geary's **c**: network autocorrelation measures, which operationalise the adage that *"birds of a feather flock together"*.

$$I = \frac{n\sum_{ij} x_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\left(\sum_{ij} x_{ij}\right)\left(\sum_i (z_i - \bar{z})^2\right)} \qquad c = \frac{(n-1)\sum_{ij} x_{ij}(z_i - z_j)^2}{2\left(\sum_{ij} x_{ij}\right)\left(\sum_i (z_i - \bar{z})^2\right)}$$

Here **z** stands for an individual variable and **x** for the network.

UCINET provides permutation-based significance levels for both statistics.

3. More conditional independence assumptions ("hybrid models")

Basic procedure:

- Calculate some meaningful measures on the actor or dyad level from the network (centrality, similarity,…)
- Treat these measures as independent variables in a "normal" (i.e., independence-assuming) statistical analysis.

Status: questionable

- It is unclear what exactly is assumed in terms of independence – formulated in general, the assumption is *"the network doesn't matter except for what we include in the analysis"* – strong danger of unobserved variable bias!
- There usually is no guiding principle that would steer the primary step of network data reduction.

4. Explicit network (& dependence) modelling

Stochastic model = model with a random component

› any stochastic model can be used to simulate many different artificial networks (a *distribution* of networks),

› by comparing simulated networks to an observed network…
- estimation of model parameters & std.errors becomes possible,
- hypothesis tests can be done based on these estimates,
- model fit can be checked on dimensions other than those included in the model.

› by comparing network distributions from different models among each other, the *interdependencies* of network-generating patterns and processes can be studied.

Basic framework for stochastic network models:

- It is assumed that *networks are random variables* (called ***X***) with a (complex) probability distribution.
- An observed network (called ***x***) is assumed to be drawn from the space of all possible networks according to this distribution.

The distribution…

- …can be formulated in a model,
- …can (at least) be simulated ("Markov Chain Monte Carlo"),
- …can be used for hypothesis testing.

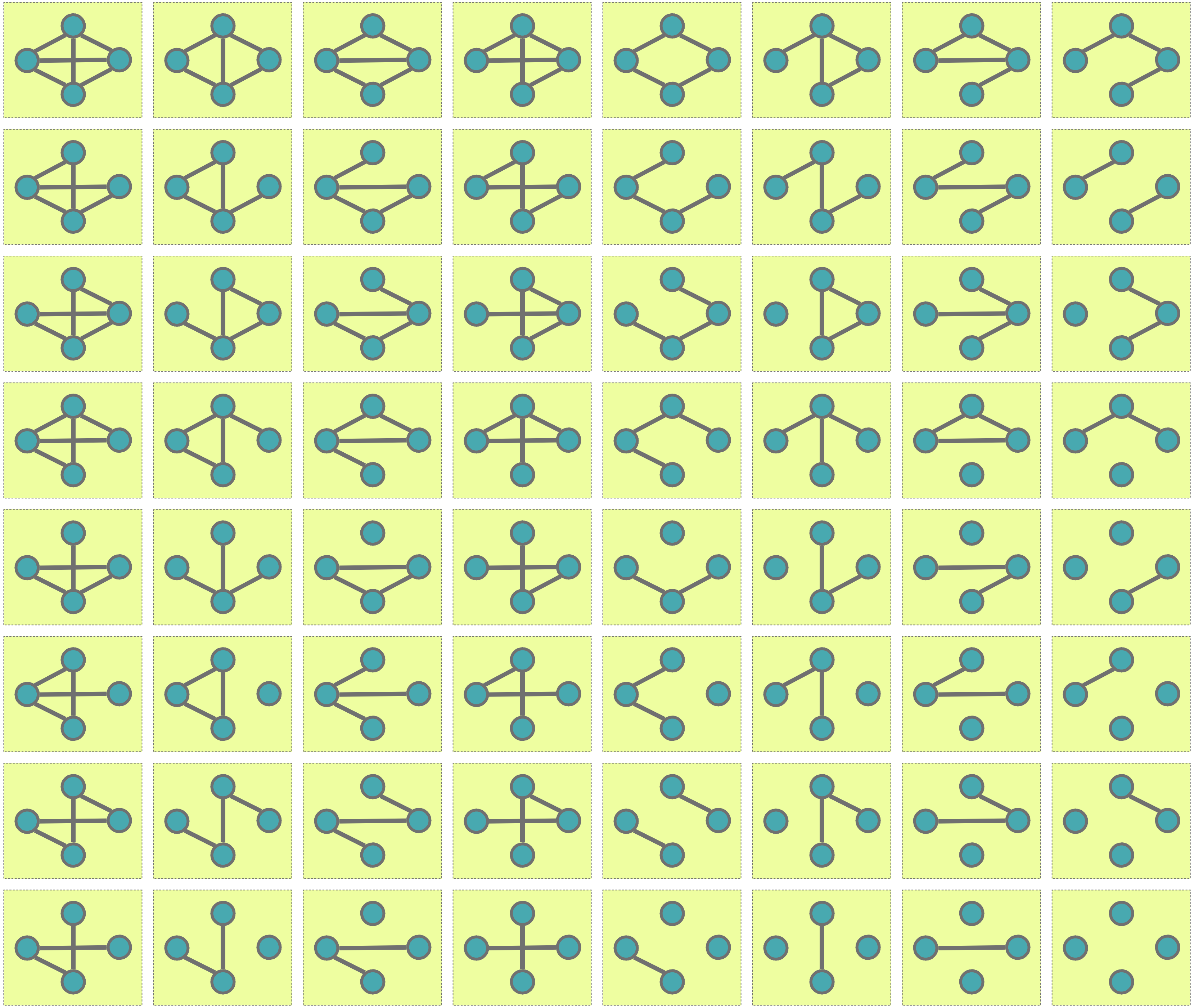
The network space is huge…

› For an undirected, binary ("zero-one") network among n actors, how many networks are possible?

- For each dyad (i,j), there are **2** possibilities: $x_{ij}=0$ or $x_{ij}=1$,
- There are $n \times (n-1)/2$ dyads ,
- Dyad outcomes can be combined in any way: totality of $2^{n \times (n-1)/2}$.

n	1	2	3	4	5	…	10	…
# of networks	1	2	8	64	1024	…	~35 trillion	…

State space for undirected networks with $\boldsymbol{n}$=**4** actors



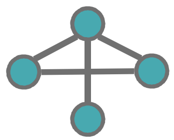
Once more independence:
The Erdős-Rényi (Bernoulli graph) model:

- Suppose all dyads are *independent*, and that a dyad $(\boldsymbol{i}, \boldsymbol{j})$ is connected with the probability $\boldsymbol{p}$.
- Then the probability of any network $\boldsymbol{x}$ can be written as the product of the dyad probabilities (simple product rule holds for independent events).
- *Formally, we have* $\Pr(\boldsymbol{X}=\boldsymbol{x}) = \mathrm{p}^{\#\mathrm{ties}} \times (\mathbf{1}-\boldsymbol{p})^{\#\mathrm{non\text{-}ties}}$,
  where $\#\text{non-ties} = (\, \boldsymbol{n} \times (\boldsymbol{n}-\mathbf{1})/\mathbf{2}) - \#\text{ties}$

The probability distribution on the network space

- …depends not on "structure" but only on tie counts!
  (see following slide)

› Now suppose that in a data collection, we observed the following particular network:



› Then the empirical tie probability is:

$$\boldsymbol{p} = \#\text{ties}/(\boldsymbol{n}\times(\boldsymbol{n}-\mathbf{1})/2) = \mathbf{2/3}$$

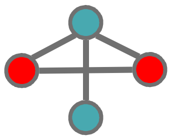
The 'best-fitting' probability distribution on the network space is given on the following slide … and has some problems:

- *Observed network is "lumped together" with other, non-equivalent networks,*
- *Highest probability has the full network, not the observed one…*

Probabilities under independence model with p=2/3

What about permutation-based distributions?

- Suppose again that in a data collection, we observed the same network:



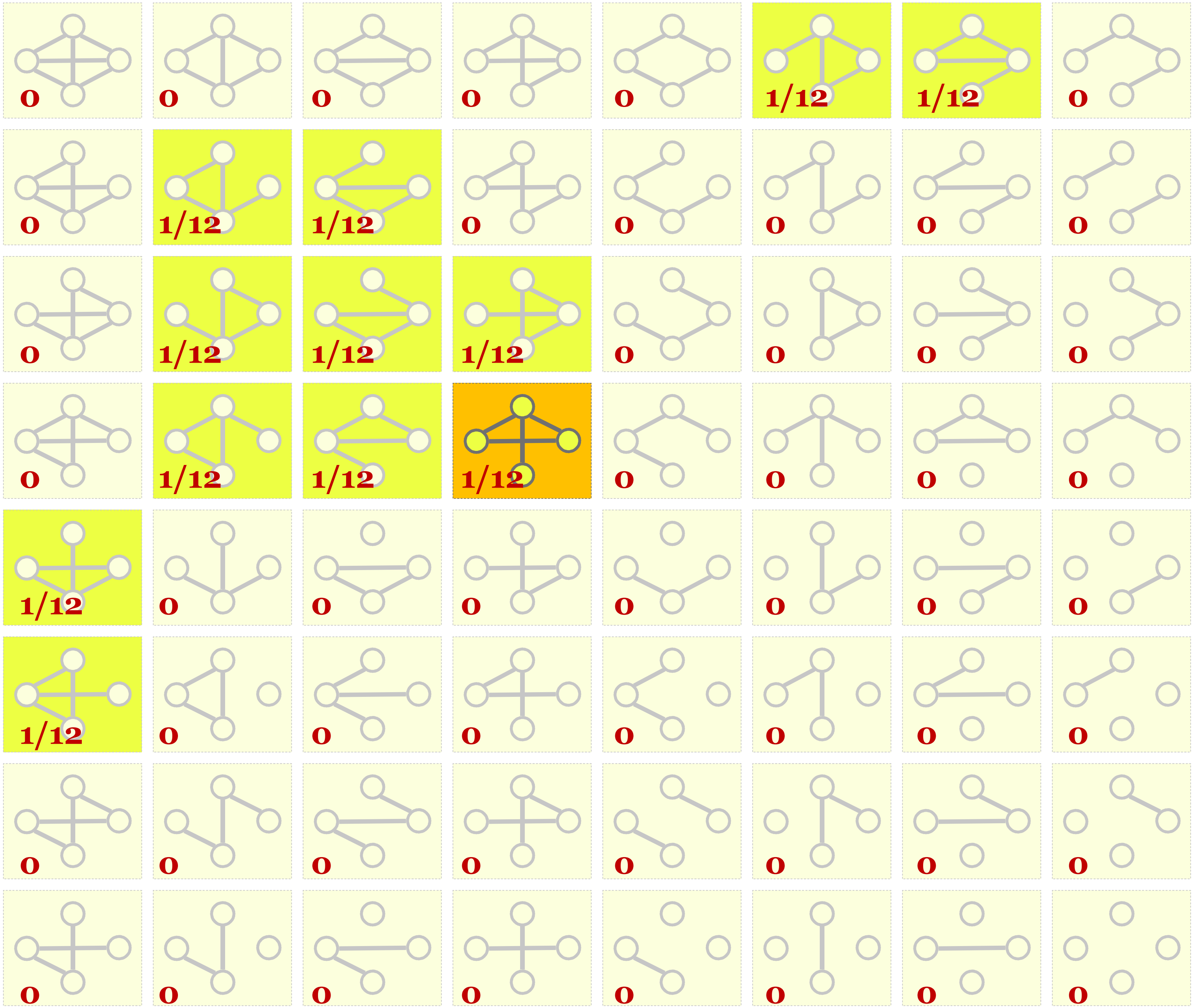
- For n=**4** actors, the number of permutations of these actors is **4!=4×3×2×1 = 24** , so there are **24** permuted networks

  …of which each has a *structurally indistinguishable twin* because the actors marked **red** above are in fully equivalent positions,

  …so **12** networks remain, they all have the same probability

  Pr(X=x) =**1/12** ≈ **8.3%** while all other networks have Pr(X=x)=**0** .
- See next slide for how the best-fitting permutation-based distribution for this network looks like.

Probabilities under permutations of the actors



What to conclude for permutation-based distributions?

- They distinguish optimally between equivalent and non-equivalent structures ("isomorphic networks"),
- and do so better than the Bernoulli graph model (4-cycles are not treated identically to the example network),
- but do *only* this and nothing else – probabilities are zero for *all* non-isomorphic networks!
- This is a bad approach when considering *measurement error:*
    - small deviations between two networks are treated the same as huge differences! Error is *inflated* this way.

Better would be a model where similar networks have similar probabilities…

Solution: explicit mathematical formulation of network probabilities

Take a "parametric approach":

specify the probability for any observed network as a mathematical function

- p1 model, p2 model, exp. random graph model
- stochastic actor-based model for network evolution

The latter treats the network state space as the state space on which a stochastic process occurs.

→ *More detail in the next series of slides.*