



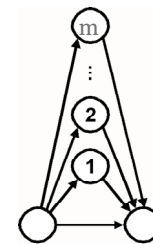
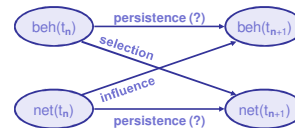
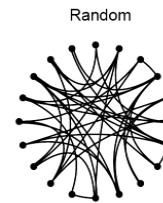
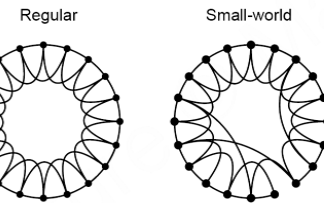
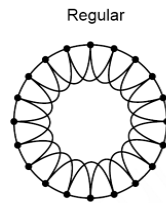
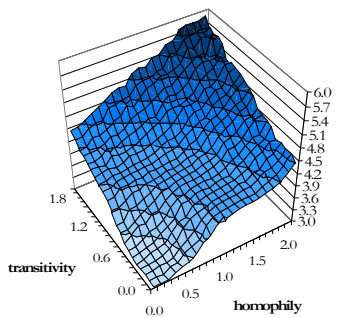
Stochastic Modelling of Networks

LINKS workshop 2009, University of Kentucky

c.e.g.steglich@rug.nl



median geodesic distance between groups



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





Overview course module “Stochastic Modelling”

I. Introduction

- A. simulation & estimation purposes of modelling
- B. testing hypotheses on network data
- C. basics of stochastic modelling

II. Actor-based models for network evolution

III. Co-evolution models for networks and actor properties

IV. Exponential Random Graph Models



A. Uses of modelling in network-related research

(1) *Simulation studies*

- Network structures as basis for secondary research,
- Aim: Generation of reasonably realistic networks.

Examples: Small worlds / grids of cellular automata / scale free networks as structure on which contagion processes are studied.

(2) *Empirical studies*

- Fitting of data sets,
- Aim: Testing of hypotheses.

The two perspectives complement each other.



(1) Models for simulation purposes

Networks are often used as “substrate” for other research:

- › epidemiology,
- › ecosystem dynamics,
- › segregation studies,
- › opinion dynamics,
- › etc.

Crucially important:

- › these ‘substrate networks’ need to mimic real networks,
- › but real networks are ill (i.e., incompletely) understood!

Task: Assess patterns & processes of real-life networks!



Which network properties to mimic?

Empirical research identified a host of network patterns

- › clustering in affective networks (friendship, trust, support),
- › brokerage in instrumental networks (professional advice, trade),
- › core-periphery structures in communities (communication),
- › homophily, balance, etc., etc.

Only a select few gained a lot of attention:

- › “small worlds” (low median geod. distance, low density, high clustering)
- › “scale-free networks” (log-degree distribution log-linear)
- › “lattice structures” (local neighbourhoods, geography-based)

Typically, simulation studies rely on these ‘select few’...



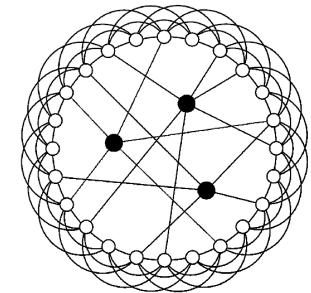
Problems with extant network simulation studies

- › The constructions are highly contingent on the specific network-generating algorithm employed.
- › The generated networks display additional features that may or may not be of merit, depending on the research question and real-life networks under study
 - e.g., small world models cannot account for clustering.

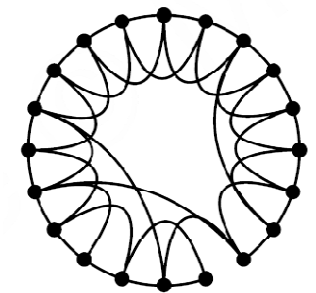
Criteria for a 'better' approach:

- › More scope in what can be modelled substantively,
- › less contamination by algorithmic artifacts.

Why not let the model 'pick up' relevant features from empirical data?



*Small worlds by
 Watts & Strogatz
 (bottom) and
 Newman (top).*





(2) Modelling of empirical data

Describing

“CSI approach”

- › Empirical data are dissected into a collection of relevant features / evidence.
- › The features are easier to understand / link up with theory.

Theory motivates / steers feature collection; validation through consistency with expectations.

Modelling

“Frankenstein approach”

- › Injection of relevant features into artificial data through model parametrisation.
- › Fine-tuning of artificial data to empirical data yields evidence for features.

Model instantiates theory; validation through goodness-of-fit tests and hypothesis tests.



B. 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*”...



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?

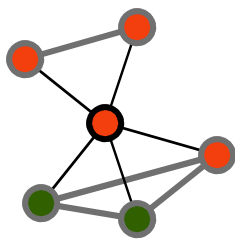
Such questions can easily be phrased in terms of hypotheses, to be tested on network data.



Complete vs. ego-centered network studies

Ego-centered data:

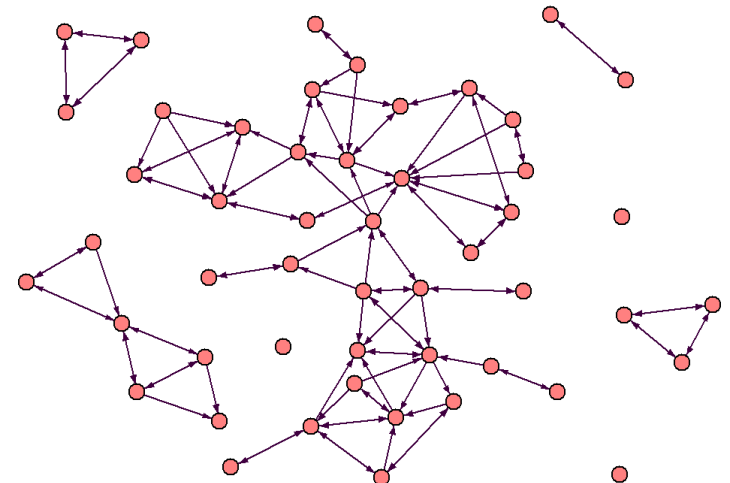
- › A random sample of actors is drawn,
- › each of the actors' network neighbourhood is measured.



*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:
 - 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!



Alternatives to sampling distributions

“Non-parametric” alternatives

1. Distributions assuming dyad independence or tie independence, given observed marginals,
2. Distributions under permutations of the actors.

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.**



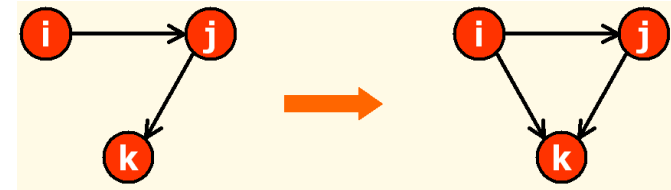
1) Distributions assuming tie/dyad independence

Exemplary question:

Is there evidence for transitive closure in a given network?

...could be operationalised

1. by counting transitive triplets (configuration on the right), 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

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

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

2. Expected fraction of transitive triplets among both configurations is

$$E(f_T) = p$$

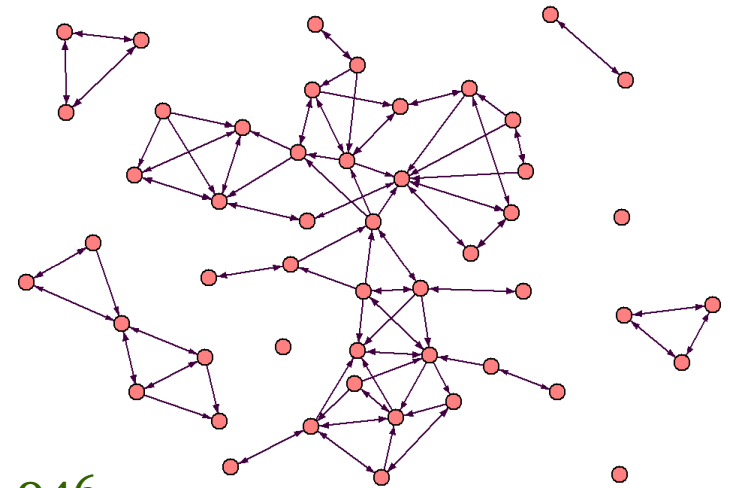
distribution is binomial $B(n(n-1), p)/(n(n-1))$.



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'

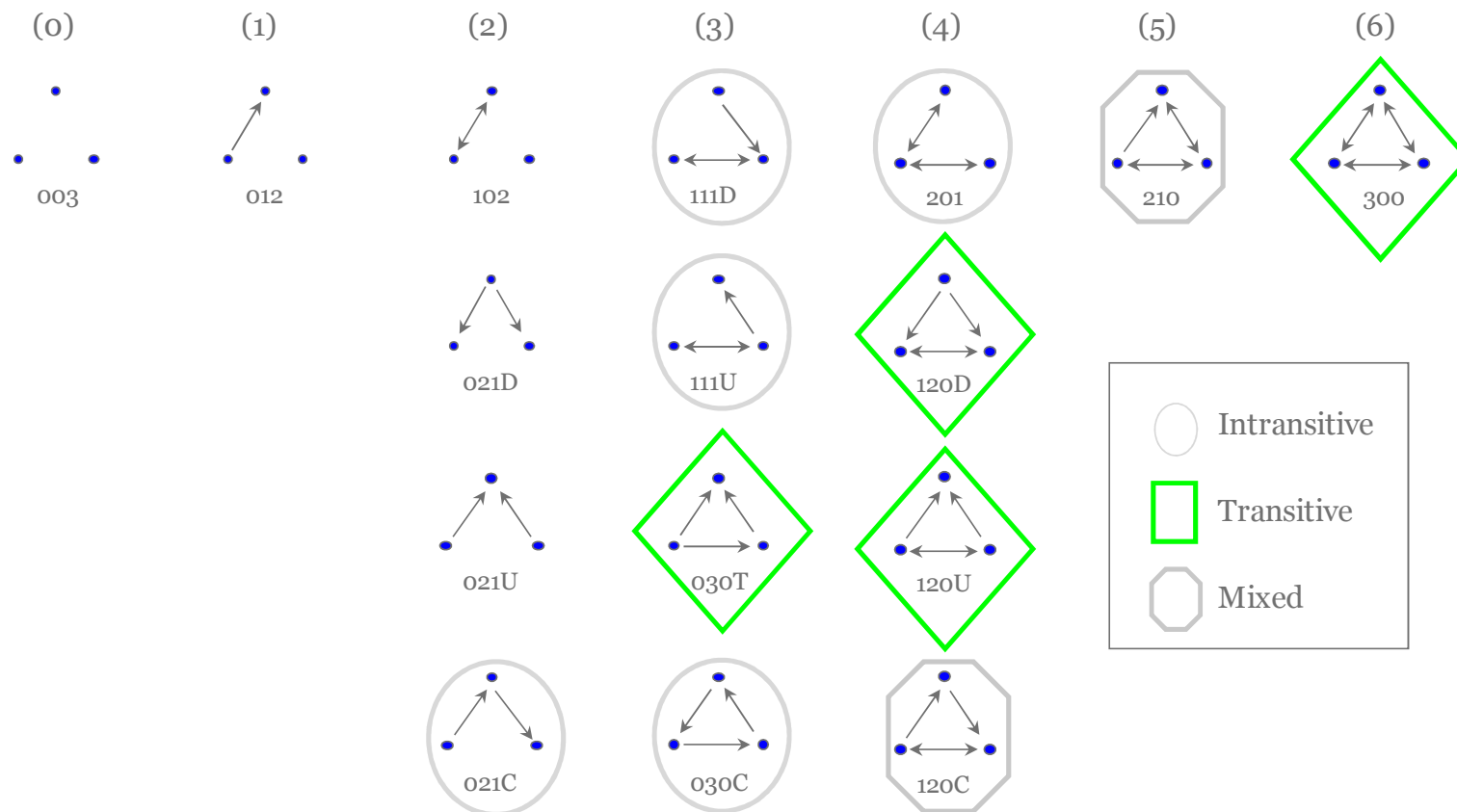
1. Expected count of transitive triplets is ~ 11.25 , p-value is far below 0.0001.
2. Expected fraction of transitive triplets is 0.046, observed fraction is $86/222=0.387$, p-value again is far below 0.0001.



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



Such a test can also be done for the whole **triad census** of the network (Holland & Leinhardt 1978):





Output (obtained with the Pajek software) :

Triadic Census 2.

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

Chi-Square: 164896.9327***

7 cells (43.75%) have expected frequencies less than 5.
The minimum expected cell frequency is 0.00.

**independence hypothesis
rejected**



Alternative, *slightly* more realistic null distributions developed in the same tradition control not just for *density* (or, equivalently, *average degree*), but for the total *degree distribution* – e.g., Karlberg (1999) proposes transitivity tests of this sort.

Problem with the whole approach: When testing structural properties, the null hypothesis of tie / dyad independence is pretty much always rejected.

So why continue to work with it at all?

...one reason may be to convince network-reluctant journal 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 (ϕ) coefficient, taking the $n(n-1)$ cells in the adjacency matrices as units of analysis.



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 >= 0	Prop <= 0
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*".

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

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

UCINET provides permutation-based significance levels for both statistics.



3. Assuming conditional independence

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*

Network models differ on many fronts. Some are these...

linear \leftrightarrow non-linear
dyad-based \leftrightarrow actor-based
static \leftrightarrow dynamic

Biased nets: *linear, dyad-based, static*

MRQAP: *linear, dyad-based, static*

ERGMs: *non-linear, dyad-based, static*

“SIENA”: *non-linear, actor-based, dynamic*

} our course
topics

Random components of these models allow hypothesis tests!



Stochastic network modelling and simulation

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 \mathbf{X}) with a (complex) probability distribution.
- › An observed network (called \mathbf{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 be simulated (“Markov Chain Monte Carlo”),
- › ...can be used for hypothesis testing.

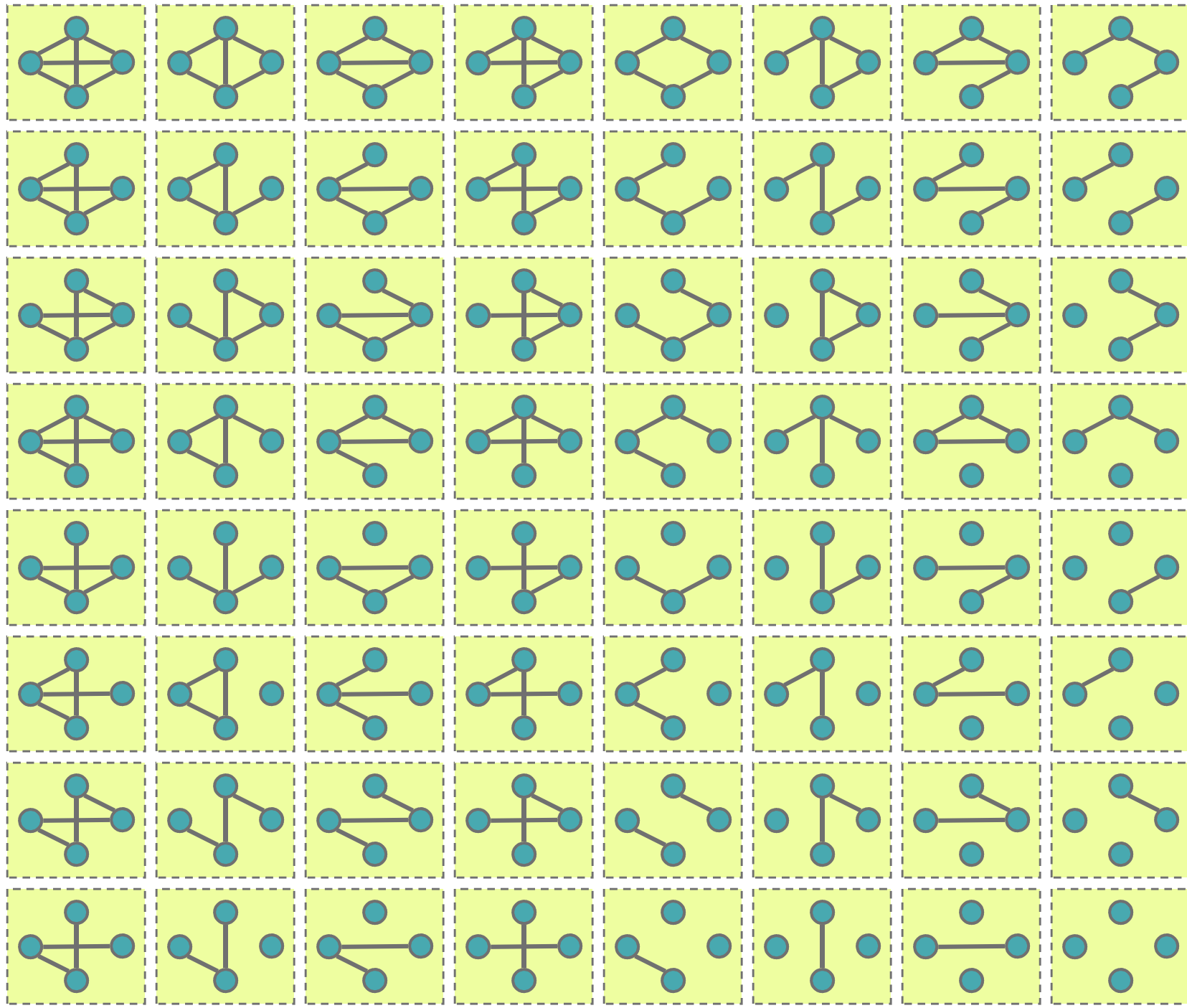


The network space is huge...

- › 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	125	...	~35 trillion	...

State space for undirected networks with $n=4$ actors





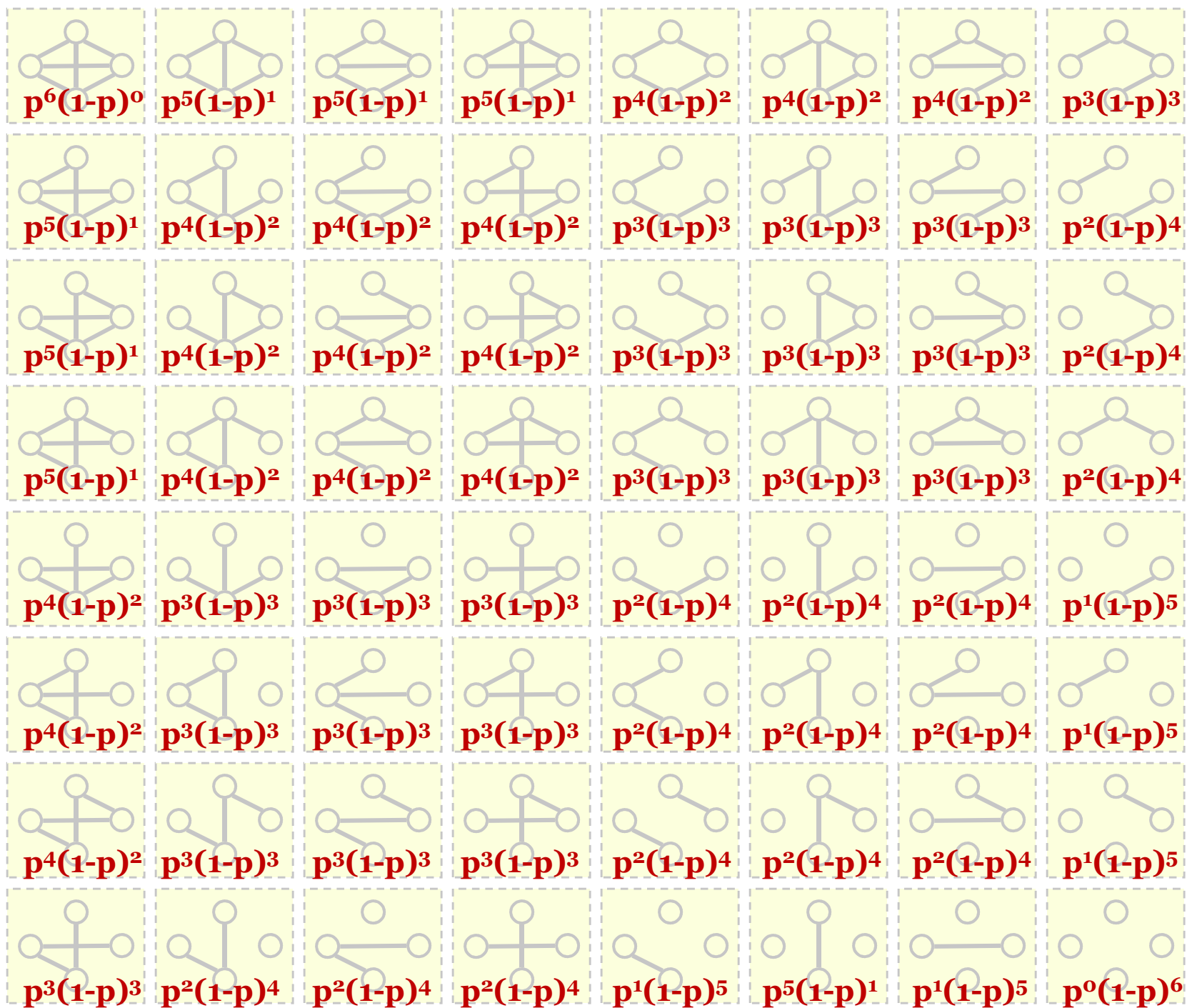
The Erdős-Rényi (Bernoulli graph) model:

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

The probability distribution on the network space

- › ...depends not on “structure” but only on tie counts!
(see following slide)

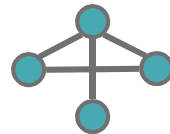
Probabilities under the Bernoulli graph model





The Erdős-Rényi (Bernoulli graph) model:

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



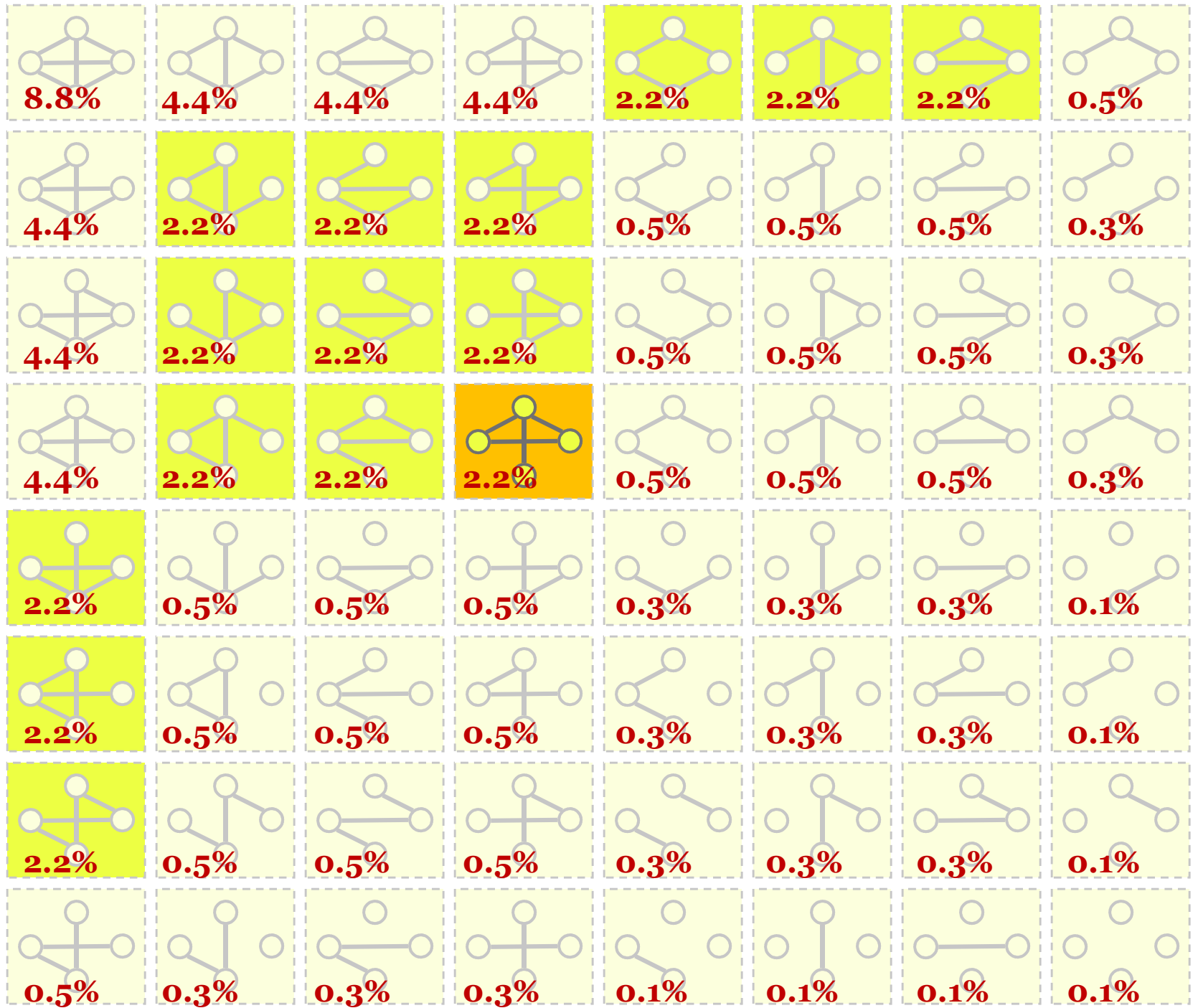
- › Then the empirical tie probability is:

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

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...*

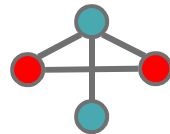
Probabilities under independence model with $p=2/3$





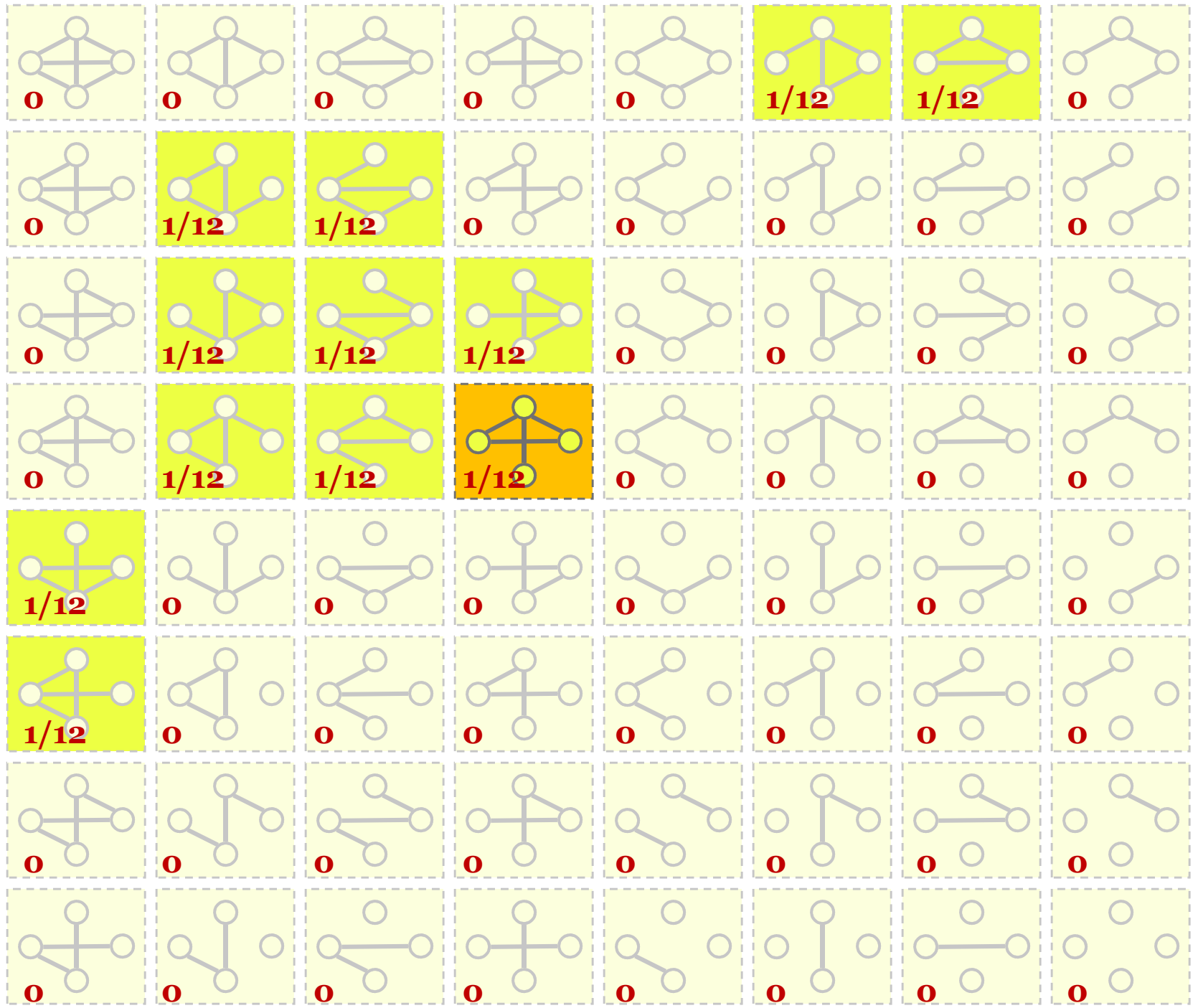
What about permutation-based distributions?

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



- › For $n=4$ actors, the number of permutations of these actors is $4! = 4 \times 3 \times 2 \times 1 = 24$, 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
 $\Pr(\mathbf{X}=\mathbf{x}) = 1/12 \approx 8.3\%$ while all other networks have $\Pr(\mathbf{X}=\mathbf{x})=0$.
- › See next slide for how the best-fitting permutation-based distribution for this network looks like.

Probabilities under permutations of the actors





What about 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... like ERGMs! → last course day.



Network dynamics necessitate a slightly different approach

- › Also to be seen as probabilistic models, on the same space of all possible networks,
- › but now we are not interested in total probability distribution over this space,
- › instead, we are interested in transition probabilities between [neighbouring] states...
- › The state space for a change process consists of all possible trajectories in our known state space, linking two networks. Still bigger!