Random Bipartite Networks

Last updated: 2023/08/22, 11:48:21 EDT

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 6701, 6713, & a pretend number,
2023–2024 | @pocsvox

Prof. Peter Sheridan Dodds | @peterdodds

Computational Story Lab | Vermont Complex Systems Center
Santa Fe Institute | University of Vermont

Outline

"Flavor network and the principles of food pairing"
Ahn et al.,
Nature Scientific Reports, **1**, 196, 2011. [1]

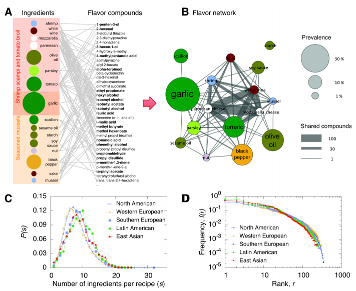


Figure 1 | Flavor network. (A) The ingredients contained in two recipes (left column), together with the flavor compounds that are known to be present in the ingredients (right column). Each flavor compound is linked to the ingredients that contain it, forming a bipartite network. Some compounds (shown in boldface) are shared by multiple ingredients. (B) If we project the ingredient-compound bipartite network into the ingredient space, we obtain the *flavor network*, whose nodes are ingredients, linked if they share at least one flavor compound. The thickness of links represents the number of flavor compounds two ingredients share and the size of each circle corresponds to the prevalence of the ingredients in recipes. (C) The distribution of recipe size, capturing the number of ingredients per recipe, across the five cuisines explored in our study. (D) The frequency-rank plot of ingredients across the five cuisines show an approximately invariant distribution across cuisines.

"Flavor network and the principles of food pairing"
Ahn et al.,
Nature Scientific Reports, **1**, 196, 2011. [1]

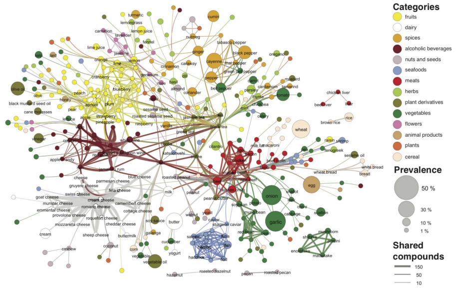


Figure 2 | The backbone of the flavor network. Each node denotes an ingredient, the node color indicates food category, and node size reflects the ingredient prevalence in recipes. Two ingredients are connected if they share a significant number of flavor compounds, link thickness representing the number of shared compounds between the two ingredients. Adjacent links are bundled to reduce the clutter. Note that the map shows only the statistically significant links, as identified by the algorithm of Refs.[28,29] for p-value 0.04. A drawing of the full network is too dense to be informative. We use, however, the full network in our subsequent measurements.

"Recipe recommendation using ingredient networks"
Teng, Lin, and Adamic,
Proceedings of the 3rd Annual ACM Web Science Conference, **1**, 298–307, 2012. [8]

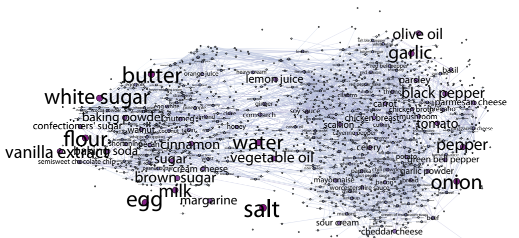
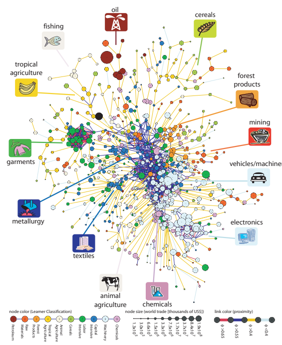


Figure 2: Ingredient complement network. Two ingredients share an edge if they occur together more than would be expected by chance and if their pointwise mutual information exceeds a threshold.

"The Product Space Conditions the Development of Nations"
Hidalgo et al.,
Science, **317**, 482–487, 2007. [6]



Networks and creativity:

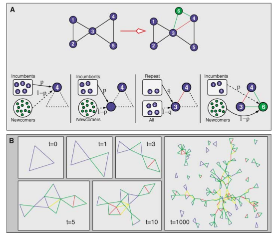
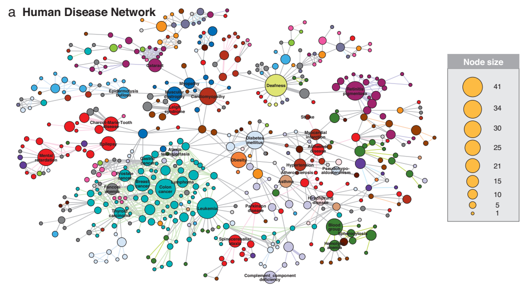


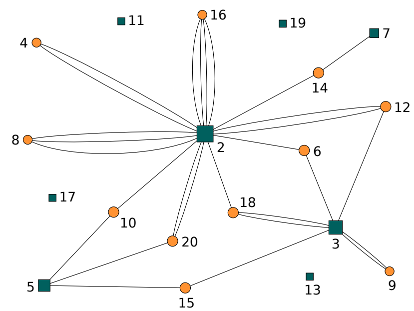
Fig. 2. Modeling the emergence of collaboration networks in creative enterprises. (**A**) Creation of a team with $m = 3$ agents. Consider, at time zero, a collaboration network comprising five agents, all incumbents (blue circles). Along with the incumbents, there is a large pool of newcomers (green circles) available to participate in new teams. Each agent in a team has a probability p of being drawn from the pool of incumbents and a probability $1 - p$ of being drawn from the pool of newcomers. For the second and subsequent agents selected from the incumbents' pool: (i) with probability q, the new agent is randomly selected from among the set of collaborators of a randomly selected incumbent already in the team; (ii) otherwise, he or she is selected at random among all incumbents in the network. For concreteness, let us assume that incumbent 4 is selected as the first agent in the new team (leftmost box). Let us also assume that the second agent is an incumbent, too (center-left box). In this example, the second agent is a past collaborator of agent 4, specifically agent 3 (center-right box). Lastly, the third agent is selected from the pool of newcomers; this agent becomes incumbent 6 (rightmost box). In these boxes and in the following panels and figures, blue lines indicate newcomer-newcomer collaborations, green lines indicate newcomer-incumbent collaborations, yellow lines indicate new incumbent-incumbent collaborations, and red lines indicate repeat collaborations. (**B**) Time evolution of the network of collaborations according to the model for $p = 0.5$, $q = 0.5$, and $m = 3$.

- Guimerà et al., Science 2005: [5] "Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance"
- Broadway musical industry
- Scientific collaboration in Social Psychology, Economics, Ecology, and Astronomy.

"The human disease network"
Goh et al.,
Proc. Natl. Acad. Sci., **104**, 8685–8690, 2007. [4]



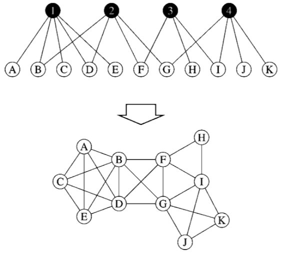
"The complex architecture of primes and natural numbers"
García-Pérez, Serrano, and Boguñá,
https://arxiv.org/abs/1402.3612, 2014. [3]

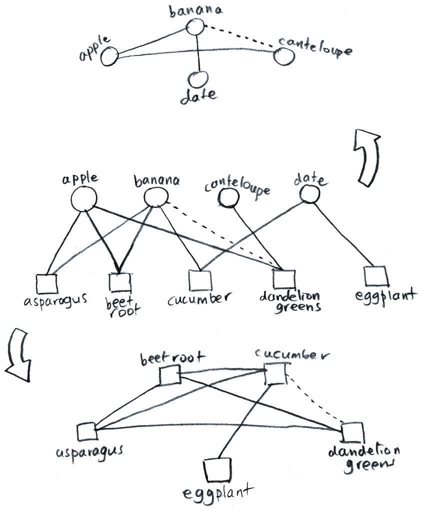


Random bipartite networks:

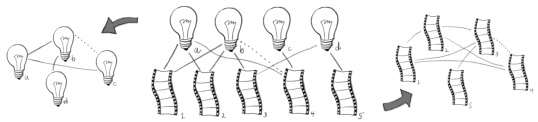
We'll follow this rather well cited paper:

"Random graphs with arbitrary degree distributions and their applications"
Newman, Strogatz, and Watts,
Phys. Rev. E, **64**, 026118, 2001. [7]





Example of a bipartite affiliation network and the induced networks:

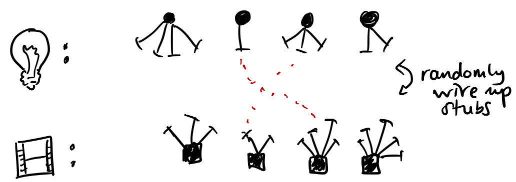


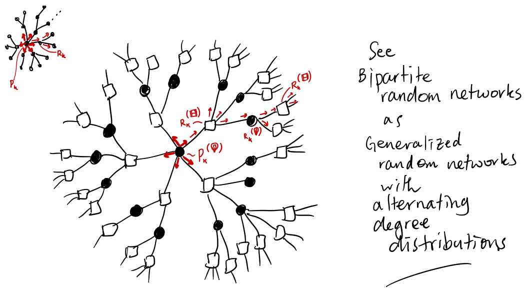
- Center: A small story-trope bipartite graph. [2]
- Induced trope network and the induced story network are on the left and right.
- The dashed edge in the bipartite affiliation network indicates an edge added to the system, resulting in the dashed edges being added to the two induced networks.

Basic story:

- An example of two inter-affiliated types:
  - 🎞 = stories,
  - 💡 = tropes.
- Stories contain tropes, tropes are in stories.
- Consider a story-trope system with $N_{🎞}$ = # stories and $N_{💡}$ = # tropes.
- $m_{🎞,💡}$ = number of edges between 🎞 and 💡.
- Let's have some underlying distributions for numbers of affiliations: $P_k^{(🎞)}$ (a story has k tropes) and $P_k^{(💡)}$ (a trope is in k stories).
- Average number of affiliations: $\langle k \rangle_{🎞}$ and $\langle k \rangle_{💡}$.
  - $\langle k \rangle_{🎞}$ = average number of tropes per story.
  - $\langle k \rangle_{💡}$ = average number of stories containing a given trope.
- Must have balance: $N_{🎞} \cdot \langle k \rangle_{🎞} = m_{🎞,💡} = N_{💡} \cdot \langle k \rangle_{💡}$.

How to build:





Usual helpers for understanding network's structure:

- Randomly select an edge connecting a 🎞 to a 💡.
- Probability the 🎞 contains k other tropes:

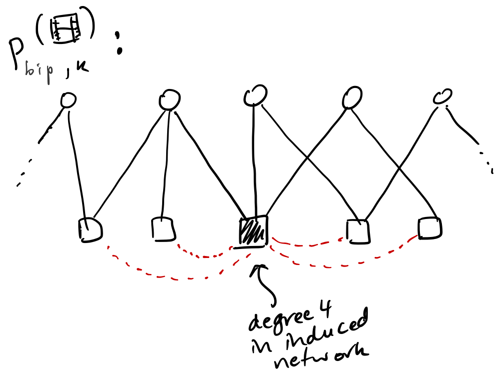
$$R_k^{(🎞)} = \frac{(k+1)P_{k+1}^{(🎞)}}{\sum_{j=0}^{N_{🎞}}(j+1)P_{j+1}^{(🎞)}} = \frac{(k+1)P_{k+1}^{(🎞)}}{\langle k \rangle_{🎞}}.$$

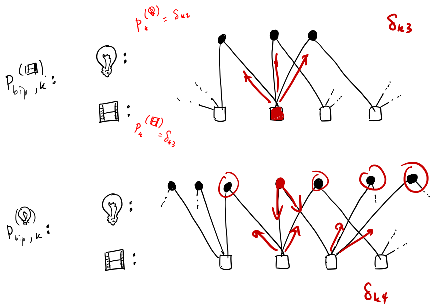
- Probability the 💡 is in k other stories:

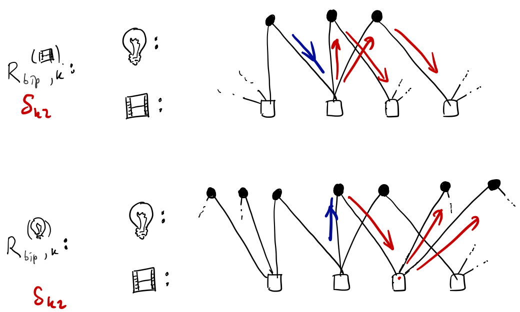
$$R_k^{(💡)} = \frac{(k+1)P_{k+1}^{(💡)}}{\sum_{j=0}^{N_{💡}}(j+1)P_{j+1}^{(💡)}} = \frac{(k+1)P_{k+1}^{(💡)}}{\langle k \rangle_{💡}}.$$

Networks of 🎞 and 💡 within bipartite structure:

- $P_{\text{ind},k}^{(🎞)}$ = probability a random 🎞 is connected to k stories by sharing at least one 💡.
- $P_{\text{ind},k}^{(💡)}$ = probability a random 💡 is connected to k tropes by co-occurring in at least one 🎞.
- $R_{\text{ind},k}^{(💡-🎞)}$ = probability a random edge leads to a 🎞 which is connected to k other stories by sharing at least one 💡.
- $R_{\text{ind},k}^{(🎞-💡)}$ = probability a random edge leads to a 💡 which is connected to k other tropes by co-occurring in at least one 🎞.
- Goal: find these distributions ☐.
- Another goal: find the induced distribution of component sizes and a test for the presence or absence of a giant component.
- Unrelated goal: be 10% happier/weep less.







Generating Function Madness

Yes, we're doing it:

- $F_{P^{(🎞)}}(x) = \sum_{k=0}^{\infty} P_k^{(🎞)} x^k$
- $F_{P^{(💡)}}(x) = \sum_{k=0}^{\infty} P_k^{(💡)} x^k$
- $F_{R^{(🎞)}}(x) = \sum_{k=0}^{\infty} R_k^{(🎞)} x^k = \frac{F'_{P^{(🎞)}}(x)}{F'_{P^{(🎞)}}(1)}$
- $F_{R^{(💡)}}(x) = \sum_{k=0}^{\infty} R_k^{(💡)} x^k = \frac{F'_{P^{(💡)}}(x)}{F'_{P^{(💡)}}(1)}$

The usual goodness:

- Normalization: $F_{P^{(🎞)}}(1) = F_{P^{(💡)}}(1) = 1$.
- Means: $F'_{P^{(🎞)}}(1) = \langle k \rangle_{🎞}$ and $F'_{P^{(💡)}}(1) = \langle k \rangle_{💡}$.

We strap these in as well:

- $F_{P_{\text{ind}}^{(🎞)}}(x) = \sum_{k=0}^{\infty} P_{\text{ind},k}^{(🎞)} x^k$
- $F_{P_{\text{ind}}^{(💡)}}(x) = \sum_{k=0}^{\infty} P_{\text{ind},k}^{(💡)} x^k$
- $F_{R_{\text{ind}}^{(💡-🎞)}}(x) = \sum_{k=0}^{\infty} R_{\text{ind},k}^{(💡-🎞)} x^k$
- $F_{R_{\text{ind}}^{(🎞-💡)}}(x) = \sum_{k=0}^{\infty} R_{\text{ind},k}^{(🎞-💡)} x^k$

So how do all these things connect?

- We're again performing sums of a randomly chosen number of randomly chosen numbers.
- We use one of our favorite sneaky tricks:

$$W = \sum_{i=1}^{U} V^{(i)} \rightleftharpoons F_W(x) = F_U(F_V(x)).$$

Induced distributions are not straightforward:

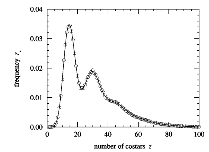
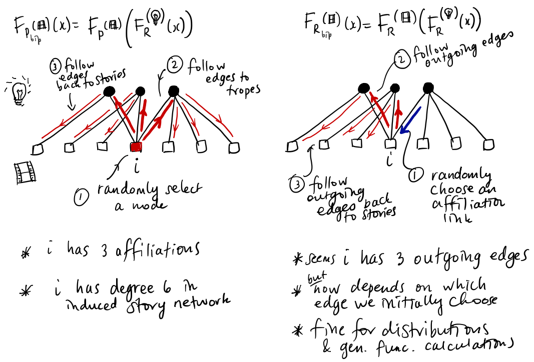
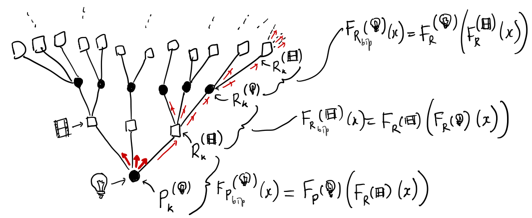


FIG. 7. The frequency distribution of numbers of co-stars of an actor in a bipartite graph with $\mu=1.5$ and $\nu=15$. The points are simulation results for $M=10\,000$ and $N=100\,000$. The line is the exact solution, Eqs. (89) and (90). The error bars on the numerical results are smaller than the points.

- View this as $P_{\text{ind},k}^{(🎞)}$ (the probability a story shares tropes with k other stories). [7]
- Result of purely random wiring with Poisson distributions for affiliation numbers.
- Parameters: $N_{🎞} = 10^4$, $N_{💡} = 10^5$, $\langle k \rangle_{🎞} = 1.5$, and $\langle k \rangle_{💡} = 15$.





Induced distribution for stories:

- Randomly choose a 🎞, find its tropes (U), and then find how many other stories each of those tropes are part of (V):

$$F_{P_{\text{ind}}^{(🎞)}}(x) = F_{P^{(🎞)}}\left(F_{R^{(💡)}}(x)\right)$$

- Find the 🎞 at the end of a randomly chosen affiliation edge leaving a trope, find its number of other tropes (U), and then find how many other stories each of those tropes are part of (V):

$$F_{R_{\text{ind}}^{(💡-🎞)}}(x) = F_{R^{(🎞)}}\left(F_{R^{(💡)}}(x)\right)$$

Induced distribution for tropes:

- Randomly choose a 💡, find the stories its part of (U), and then find how many other tropes are part of those stories (V):

$$F_{P_{\text{ind}}^{(💡)}}(x) = F_{P^{(💡)}}\left(F_{R^{(🎞)}}(x)\right)$$

- Find the 💡 at the end of a randomly chosen affiliation edge leaving a story, find the number of other stories that use it (U), and then find how many other tropes are in those stories (V):

$$F_{R_{\text{ind}}^{(🎞-💡)}}(x) = F_{R^{(💡)}}\left(F_{R^{(🎞)}}(x)\right)$$

Let's do some good:

- Average number of stories connected to a story through trope-space:

$$\langle k\rangle_{\boxplus,\mathrm{ind}} = F'_{P^{(\boxplus)}_{\mathrm{ind}}}(1)$$

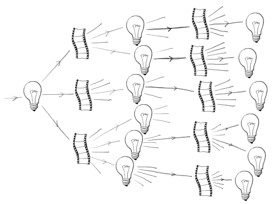
- So: $$\langle k\rangle_{\boxplus,\mathrm{ind}} = \frac{\mathrm{d}}{\mathrm{d}x} F_{P^{(\boxplus)}}\left(F_{R^{(\lozenge)}}(x)\right)\Big|_{x=1}$$
$$= F'_{R^{(\lozenge)}}(1) F'_{P^{(\boxplus)}}\left(F_{R^{(\lozenge)}}(1)\right) = F'_{R^{(\lozenge)}}(1) F'_{P^{(\boxplus)}}(1)$$

- Similarly, the average number of tropes connected to a random trope through stories:

$$\langle k\rangle_{\lozenge,\mathrm{ind}} = F'_{R^{(\boxplus)}}(1) F'_{P^{(\lozenge)}}(1)$$

- In terms of the underlying distributions, we have: $\langle k\rangle_{\boxplus,\mathrm{ind}} = \frac{\langle k(k-1)\rangle_{\lozenge}}{\langle k\rangle_{\lozenge}}\langle k\rangle_{\boxplus}$ and $\langle k\rangle_{\lozenge,\mathrm{ind}} = \frac{\langle k(k-1)\rangle_{\boxplus}}{\langle k\rangle_{\boxplus}}\langle k\rangle_{\lozenge}$

Spreading through bipartite networks:



- View as bouncing back and forth between the two connected populations. [2]
- Actual spread may be within only one population (ideas between between people) or through both (failures in physical and communication networks).
- The gain ratio for simple contagion on a bipartite random network = product of two gain ratios.

Unstoppable spreading: Is this thing connected?

- Always about the edges: when following a random edge toward a $\boxplus$, what's the expected number of new edges leading to other stories via tropes?
- We want to determine $\langle k\rangle_{R,\boxplus,\mathrm{ind}} = F'_{R^{(\lozenge-\boxplus)}_{\mathrm{ind}}}(1)$ (and $F'_{R^{(\boxplus-\lozenge)}_{\mathrm{ind}}}(1)$ for the trope side of things).
- We compute with joy:

$$\langle k\rangle_{R,\boxplus,\mathrm{ind}} = \frac{\mathrm{d}}{\mathrm{d}x} F_{R^{(\lozenge-\boxplus)}_{\mathrm{ind},k}}(x)\Big|_{x=1} = \frac{\mathrm{d}}{\mathrm{d}x} F_{R^{(\boxplus)}}\left(F_{R^{(\lozenge)}}(x)\right)\Big|_{x=1}$$

$$= F'_{R^{(\lozenge)}}(1) F'_{R^{(\boxplus)}}\left(F_{R^{(\lozenge)}}(1)\right) = F'_{R^{(\lozenge)}}(1) F'_{R^{(\boxplus)}}(1) = \frac{F''_{P^{(\lozenge)}}(1)}{F'_{P^{(\lozenge)}}(1)} \frac{F''_{P^{(\boxplus)}}(1)}{F'_{P^{(\boxplus)}}(1)}$$

- Note symmetry.
- $happiness++;

- In terms of the underlying distributions:

$$\langle k\rangle_{R,\boxplus,\mathrm{ind}} = \frac{\langle k(k-1)\rangle_{\boxplus}}{\langle k\rangle_{\boxplus}} \frac{\langle k(k-1)\rangle_{\lozenge}}{\langle k\rangle_{\lozenge}}$$

- We have a giant component in both induced networks when
$$\langle k\rangle_{R,\boxplus,\mathrm{ind}} \equiv \langle k\rangle_{R,\lozenge,\mathrm{ind}} > 1$$
.
- See this as the product of two gain ratios. #excellent #physics
- We can mess with this condition to make it mathematically pleasant and pleasantly inscrutable:

$$\sum_{k=0}^{\infty}\sum_{k'=0}^{\infty} kk'(kk'-k-k')P^{(\boxplus)}_k P^{(\lozenge)}_{k'} = 0.$$

Simple example for finding the degree distributions for the two induced networks in a random bipartite affiliation structure:

- Set $P^{(\boxplus)}_k = \delta_{k3}$ and leave $P^{(\lozenge)}_k$ arbitrary.
- Each story contains exactly three tropes.
- We have $F_{P^{(\boxplus)}}(x) = x^3$ and $F_{R^{(\boxplus)}}(x) = x^2$.
- Using $F_{P^{(\boxplus)}_{\mathrm{ind}}}(x) = F_{P^{(\boxplus)}}\left(F_{R^{(\lozenge)}}(x)\right)$ and $F_{P^{(\lozenge)}_{\mathrm{ind}}}(x) = F_{P^{(\lozenge)}}\left(F_{R^{(\boxplus)}}(x)\right)$ we have $F_{P^{(\boxplus)}_{\mathrm{ind}}}(x) = \left[F_{R^{(\lozenge)}}(x)\right]^3$ and $F_{P^{(\lozenge)}_{\mathrm{ind}}}(x) = F_{P^{(\lozenge)}}\left(x^2\right)$.
- Even more specific: If each trope is found in exactly two stories then $F_{P^{(\lozenge)}} = x^2$ and $F_{R^{(\lozenge)}} = x$ giving $F_{P^{(\boxplus)}_{\mathrm{ind}}}(x) = x^3$ and $F_{P^{(\lozenge)}_{\mathrm{ind}}}(x) = x^4$.
- Yes for giant components ☐: $\langle k\rangle_{R,\boxplus,\mathrm{ind}} \equiv \langle k\rangle_{R,\lozenge,\mathrm{ind}} = 2\cdot 1 = 2 > 1.$

Boards and Directors: [7]

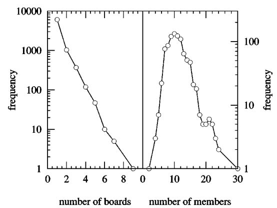


FIG. 8. Frequency distributions for the boards of directors of the Fortune 1000. Left panel: the numbers of boards on which each director sits. Right panel: the numbers of directors on each board.

- Exponentialish distribution for number of boards each director sits on.
- Boards typically have 5 to 15 directors.
- Plan: Take these distributions, presume random bipartite structure and generate co-director network and board interlock network.

Boards and Directors and more: [7]

TABLE I. Summary of results of the analysis of four collaboration networks.

Network	Clustering C Theory	Clustering C Actual	Average degree z Theory	Average degree z Actual
Company directors	0.590	0.588	14.53	14.44
Movie actors	0.084	0.199	125.6	113.4
Physics (arxiv.org)	0.192	0.452	16.74	9.27
Biomedicine (MEDLINE)	0.042	0.088	18.02	16.93

- Random bipartite affiliation network assumption produces decent matches for some basic quantities.

Boards and Directors: [7]

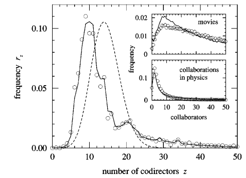


FIG. 9. The probability distribution of numbers of co-directors in the Fortune 1000 graph. The points are the real-world data, the solid line is the bipartite graph model, and the dashed line is the Poisson distribution with the same mean. Insets: the equivalent distributions for the numbers of collaborators of movie actors and physicists.

- Jolly good: Works very well for co-directors.
- For comparison, the dashed line is a Poisson with the empirical average degree.

Boards and Directors: [7]

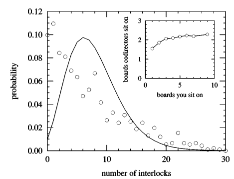


FIG. 10. The distribution of the number of other boards with which each board of directors is "interlocked" in the Fortune 1000 data. An interlock between two boards means that they share one or more common members. The points are the empirical data, the solid line is the theoretical prediction. Inset: the number of boards on which one's codirectors sit, as a function of the number of boards one sits on oneself.

- Wins less bananas for the board interlock network.
- Assortativity is the reason: Directors who sit on many boards tend to sit on the same boards.
- Note: The term assortativity was not used in this 2001 paper.

To come:

- Distributions of component size.
- Simpler computation for the giant component condition.
- Contagion.
- Testing real bipartite structures for departure from randomness.

Nutshell:

- Random bipartite networks model many real systems well.
- Crucial improvement over simple random networks.
- We can find the induced distributions and determine connectivity/contagion condition.

References I

[1] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási.
Flavor network and the principles of food pairing.
Nature Scientific Reports, 1:196, 2011. pdf

[2] P. S. Dodds.
A simple person's approach to understanding the contagion condition for spreading processes on generalized random networks.
In S. Lehmann and Y.-Y. Ahn, editors, Spreading Dynamics in Social Systems. 2017. pdf

[3] L. P. García-Pérez, M. A. Serrano, and M. Boguñá.
The complex architecture of primes and natural numbers, 2014.
https://arxiv.org/abs/1402.3612. pdf

References II

[4] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási.
The human disease network.
Proc. Natl. Acad. Sci., 104:8685–8690, 2007. pdf

[5] R. Guimerà, B. Uzzi, J. Spiro, and L. A. N. Amaral.
Team assembly mechanisms determine collaboration network structure and team performance.
Science, 308:697–702, 2005. pdf

[6] C. A. Hidalgo, B. Klinger, A.-L. Barabási, and R. Hausman.
The product space conditions the development of nations.
Science, 317:482–487, 2007. pdf

References III

[7] M. E. J. Newman, S. H. Strogatz, and D. J. Watts.
Random graphs with arbitrary degree distributions and their applications.
Phys. Rev. E, 64:026118, 2001. pdf

[8] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic.
Recipe recommendation using ingredient networks.
In Proceedings of the 3rd Annual ACM Web Science Conference, WebSci '12, pages 298–307, New York, NY, USA, 2012. ACM. pdf