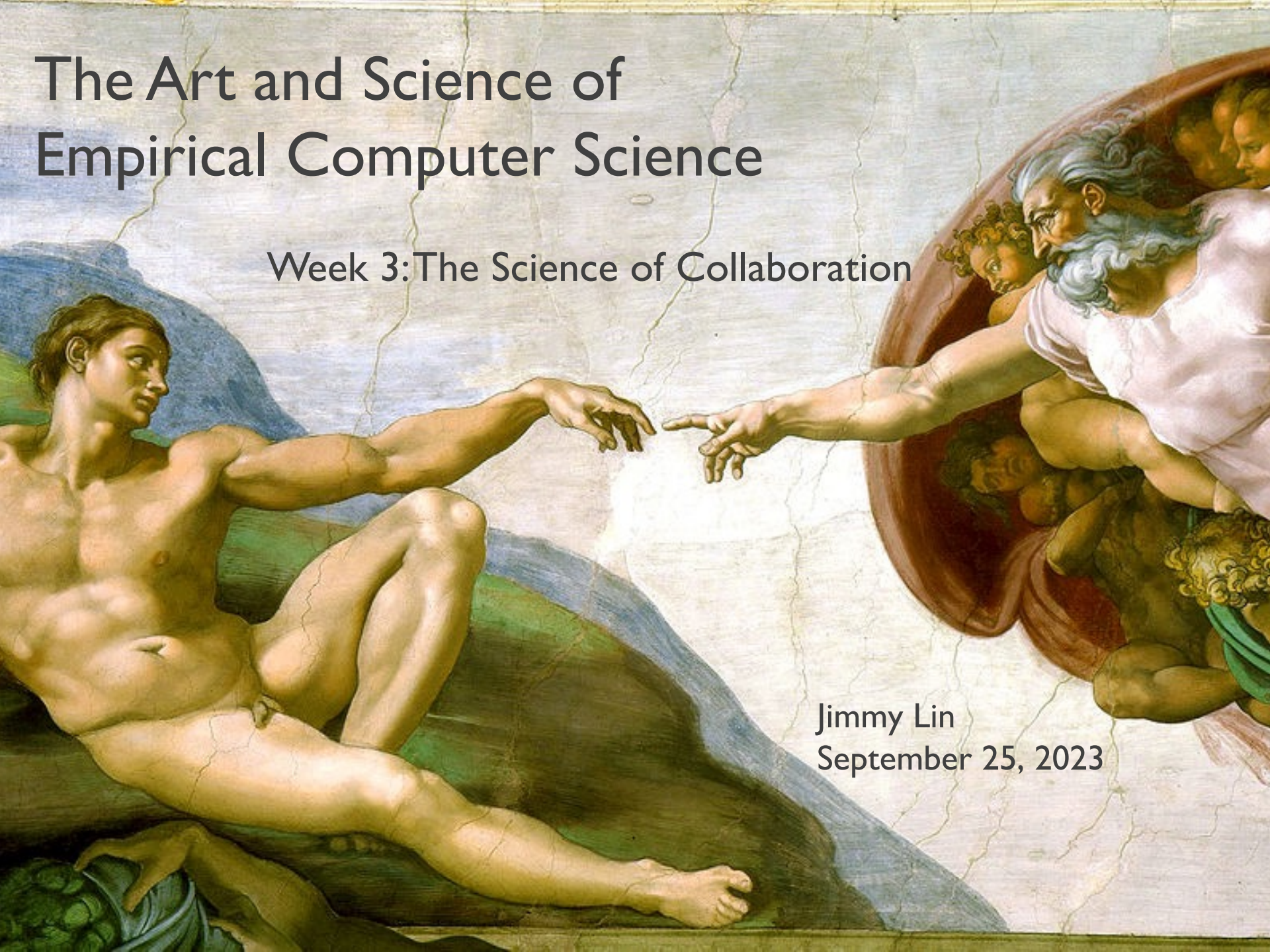
The Art and Science of Empirical Computer Science

Week 3: The Science of Collaboration



Jimmy Lin
September 25, 2023

Should you collaborate or not?

Position A: Early-stage researchers should actively seek out collaborations beyond their research group. Participation in multiple research projects across many different groups builds breadth.

Position B: Early-stage researchers should *not* actively seek out collaborations beyond their research group. Focusing on depth is more important than breadth.

Let’s debate!

"The Science of Science", Part II

The Science of Collaboration

Chapter 8: Dominance of Teams in Science

Discussion Points

Makes sense?

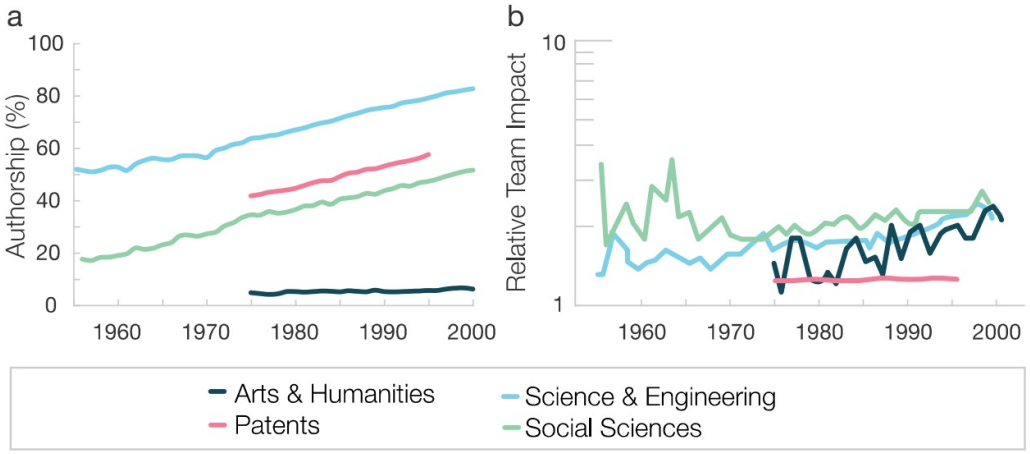


Figure 2.1.1 **The growing dominance of teams.** (a) Changes in the fraction of papers and patents written by teams over the past five decades. Each line represents the arithmetic average taken over all subfields in each year, with colors indicating different fields. (b) The Relative Team Impact (RTI) represents the mean number of citations received by team-authored work divided by the mean number of citations received by solo-authored work in the same field. A ratio of 1 implies that team- and solo-authored work have comparable impact. The lines present the arithmetic average of RTI in a given year for the entire field. After Wuchty *et al.* [2].

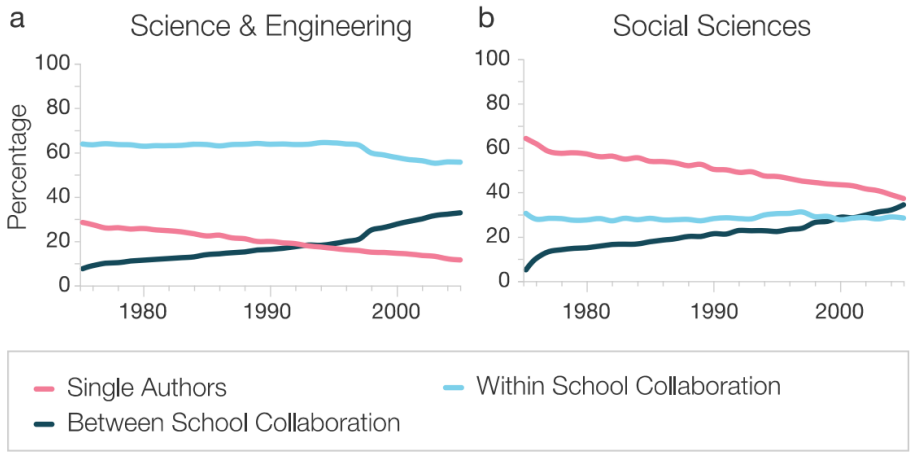


Figure 2.1.3 **The rise in multi-university collaboration**. By comparing the percentage of papers produced by different authorship combinations, the plots document the increasing share of multi-university collaborations between 1975 and 2005. This rise is especially strong in Science and Engineering (a) and Social Science (b), whereas it remains weak in Arts & Humanities, in which collaboration of any kind is rare [14]. The share of single-university collaborations remains roughly constant with time, whereas the share of solo-authored papers strongly declined in Science & Engineering and Social Sciences. After Jones *et al.* [14].

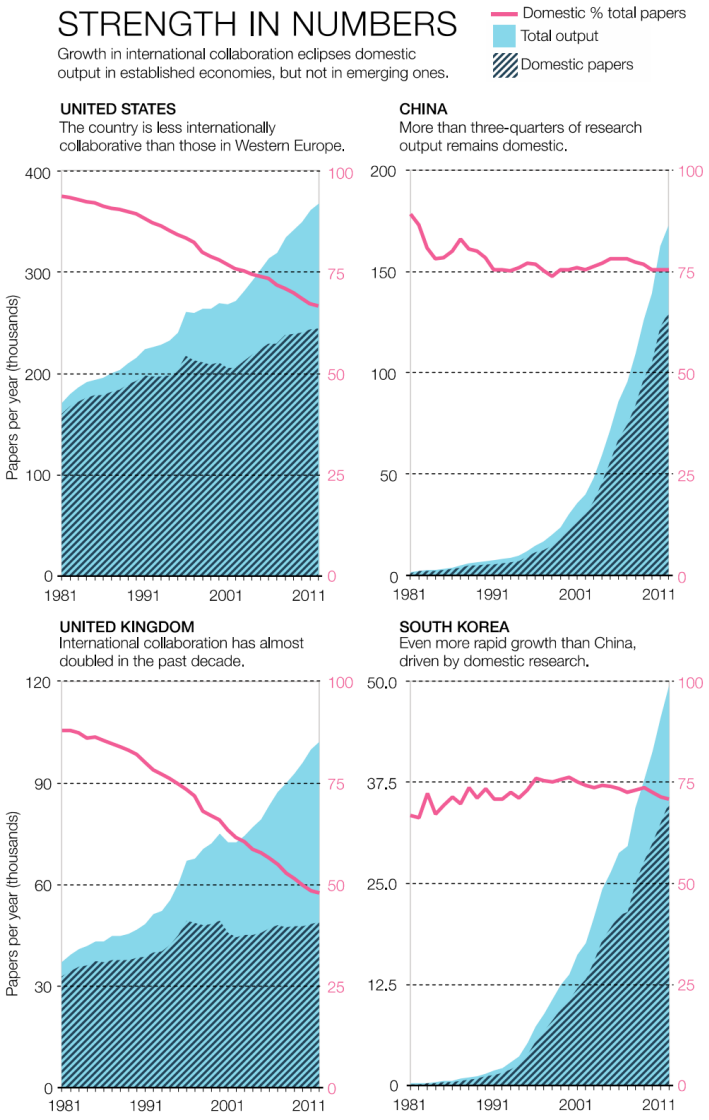


Figure 2.1.4 **The increasing role of international collaboration**. If a paper only contains authors whose addresses are from the home country, then it is counted as domestic output of that country. Comparing the left and right panels shows that growth in international collaboration eclipses the growth of domestic output in established economies, but not in emerging ones. After Adams [16].

"The Science of Science", Part II

The Science of Collaboration

Chapter 9: The Invisible College

Discussion Points

Peer effects are real, but causality is complicated…

"The Science of Science", Part II

The Science of Collaboration

Chapter 10: Coauthorship Networks

Discussion Points

Power Law distribution of vertex degrees

"Small World" phenomenon

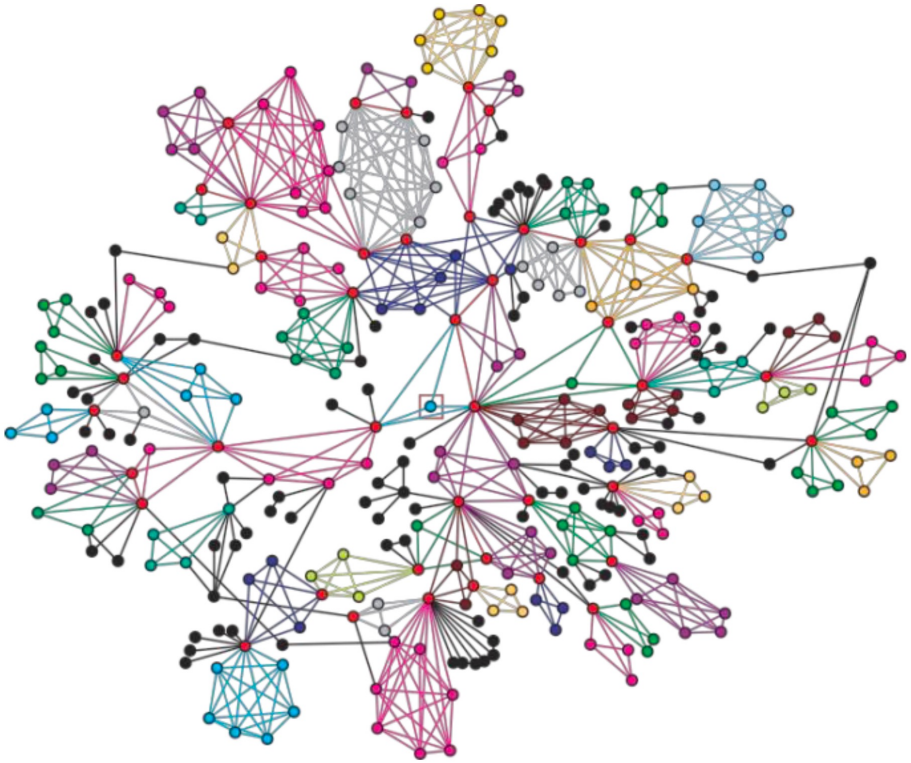


Figure 2.3.1 **Co-authorship network.** The figure shows the local structure of the co-authorship network between physicists in the vicinity of a randomly chosen individual (marked by a red frame). The network is constructed based on papers from Cornell University's archive server (cond-mat), the precursor of the widely used arXiv, containing at that time over 30,000 authors. Each node is a scientist, and links document collaborative relationships in the form of co-authored publications. The color-coded communities represent groups of collaborators that belong to locally densely interconnected parts within the network. Black nodes/edges mark those scientists that do not belong to any community. After Palla *et al.* [34].

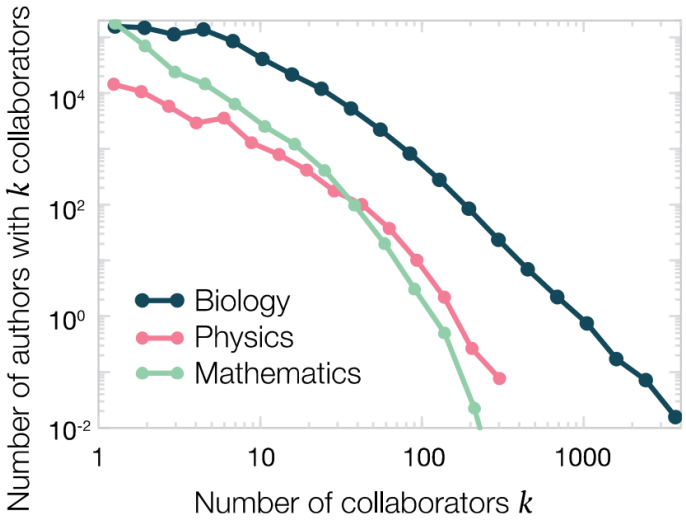


Figure 2.3.2 **Collaboration Networks are Scale-free.** The plots show the distribution of numbers of collaborators for scientists in physics, biology and mathematics, indicating that the underlying distribution are fat tailed. This implies that the collaboration network is scale-free [40, 41], meaning that the degrees can be approximated by a power law distribution. After Newman *et al.* [38].

Milgram, S. (1967) The Small World Problem.
Psychology Today.

The Small-World Problem

By Stanley Milgram

Fred Jones of Peoria, sitting in a sidewalk cafe in Tunis, and needing a light for his cigarette, asks the man at the next table for a match. They fall into conversation; the stranger is an Englishman who, it turns out, spent several months in Detroit studying the operation of an interchangeable-bottlecap-factory. "I know it's a foolish question," says Jones, "but did you ever by any chance run into a fellow named Ben Arkadian? He's an old friend of mine, manages a chain of supermarkets in Detroit . . ."

"Arkadian, Arkadian," the Englishman mutters. "Why, upon my soul, I believe I do! Small chap, very energetic, raised merry hell with the factory over a shipment of defective bottlecaps."

"No kidding!" Jones exclaims in amazement.

"Good lord, it's a small world, isn't it?'

Watts, D. J. and Strogatz, S. H. (1998). Collective Dynamics of 'Small-World' Networks. *Nature*, 393(6684): 440–442.

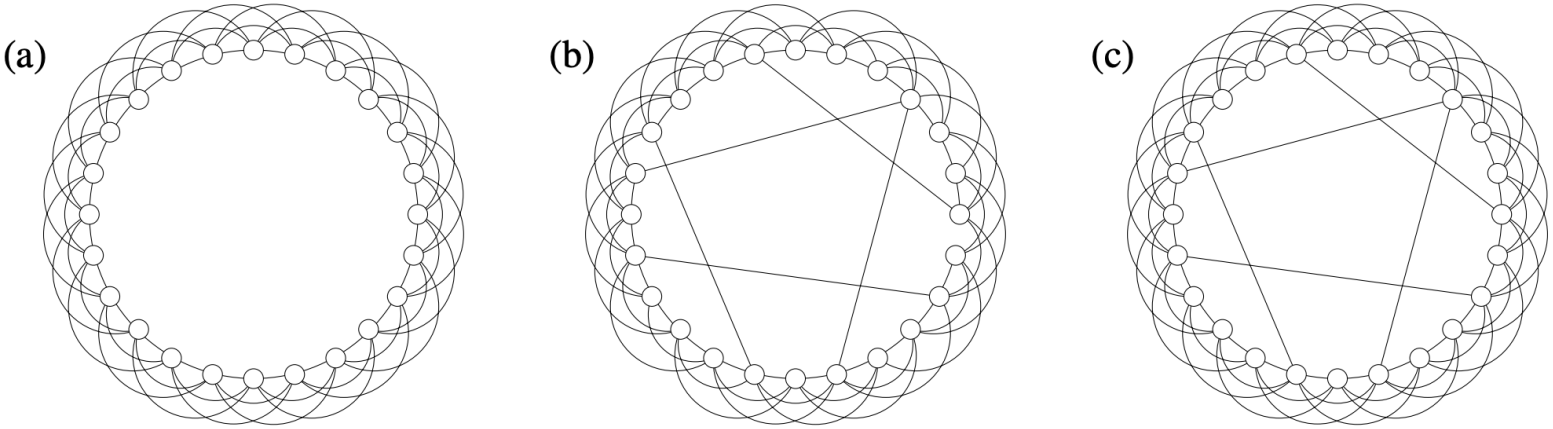


Fig. 6.1 (a) *A one-dimensional lattice with connections between all vertex pairs separated by* k *or fewer lattice spacing, with* $k = 3$ *in this case.* (b) *The small-world model* [415, 411] *is created by choosing at random a fraction* p *of the edges in the graph and moving one end of each to a new location, also chosen uniformly at random.* (c) *A slight variation on the model* [323, 288] *in which shortcuts are added randomly between vertices, but no edges are removed from the underlying one-dimensional lattice.*

From Newman (2003)

Can we say anything about these "shortcuts"?

Granovetter, M. (1973) The Strength of Weak Ties.
American Journal of Sociology, 78(6):1360-1380.

The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns in any convincing way. Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, and political structure. At the micro level, a large and increasing body of data and theory offers useful and illuminating ideas about what transpires within the confines of the small group. But how interaction in

REPORT

SOCIAL NETWORKS

A causal test of the strength of weak ties

Karthik Rajkumar[1], Guillaume Saint-Jacques[1], Iavor Bojinov[2], Erik Brynjolfsson[3,4], Sinan Aral[5]*

The authors analyzed data from multiple large-scale randomized experiments on LinkedIn's People You May Know algorithm, which recommends new connections to LinkedIn members, to test the extent to which weak ties increased job mobility in the world's largest professional social network. The experiments randomly varied the prevalence of weak ties in the networks of over 20 million people over a 5-year period, during which 2 billion new ties and 600,000 new jobs were created. The results provided experimental causal evidence supporting the strength of weak ties and suggested three revisions to the theory. First, the strength of weak ties was nonlinear. Statistical analysis found an inverted U-shaped relationship between tie strength and job transmission such that weaker ties increased job transmission but only to a point, after which there were diminishing marginal returns to tie weakness. Second, weak ties measured by interaction intensity and the number of mutual connections displayed varying effects. Moderately weak ties (measured by mutual connections) and the weakest ties (measured by interaction intensity) created the most job mobility. Third, the strength of weak ties varied by industry. Whereas weak ties increased job mobility in more digital industries, strong ties increased job mobility in less digital industries.

The Strength of Weak Ties (*1*) is one of the most influential social theories of the past century, underpinning networked theories of information diffusion (*2*, *3*), social contagion (*4*, *5*), social movements (*6*), innovation diffusion (*7*), organization (*8*), and human cooperation (*9*, *10*). It argues that infrequent, arms-length relationships (weak ties) provide disproportionately employment opportunities (*11*), promotions and greater wage increases (*12*), creativity (*13*), innovation (*14*, *15*), productivity (*16*), and performance (*17*) because they deliver more novel information than strong ties. Weak ties are thought to provide access to diverse, novel information because they connect us to disparate and diverse parts of the human social network (*18*–*24*). In addition to productivity work is not experimental the authors rightfully acknowledge that their results "may not be the true causal effect of tie strength on the probability of a sequential job." More generally, two empirical challenges have prevented robust tests of the strength of weak ties theory. First, a lack of large-scale data linking human social networks to job transmission makes measurement of the causal impact of tie strength and labor market outcomes difficult. Second, network ties and labor market outcomes are endogenous, making the causal link between weak ties and job placement elusive. Individuals' labor market outcomes are likely to be determined by and to simultaneously determine their social networks. The evolution of social networks and job trajectories are also likely correlated with unobserved factors such

invite from the ego, (iii) a model estimating the engagement between the ego and alter once connected and (iv) weights on each of these models for relative importance. The experiments tuned these components, introduced new data sources, and relied on the number of mutual connections between the ego and a potential tie recommendation as one of the most important features of the ensemble model to randomly vary weak and strong tie recommendations. We performed a retrospective analysis of the randomization created by the PYMK experiments conducted by LinkedIn between 2015 and 2019 in two waves.

The first wave examined a global experiment conducted in 2015 that had over 4 million experimental subjects and created over 19 million new connections. We collected edge-level observations of tie strength and job transmission outcomes for each tie created during this experiment. We then analyzed a larger second wave of node-level PYMK experiments that took place worldwide in 2019. The second wave spanned every continent and US state, had more than 16 million experimental subjects, created ~2 billion new connections and recorded more than 70 million job applications that led to 600,000 new jobs during the experimental period (Fig. 1, B and C). The data [illegible] where each observation corresponds to a unique LinkedIn member, and at the edge level, [illegible] each observation corresponds to a unique tie between two LinkedIn members [illegible] tion of how we compiled the edge- and node-level datasets).

We analyzed labor market mobility by measuring both job applications and job transmissions. Job applications are simply the number of jobs LinkedIn members applied to on the platform in the three months after an experiment. In accordance with the litera-

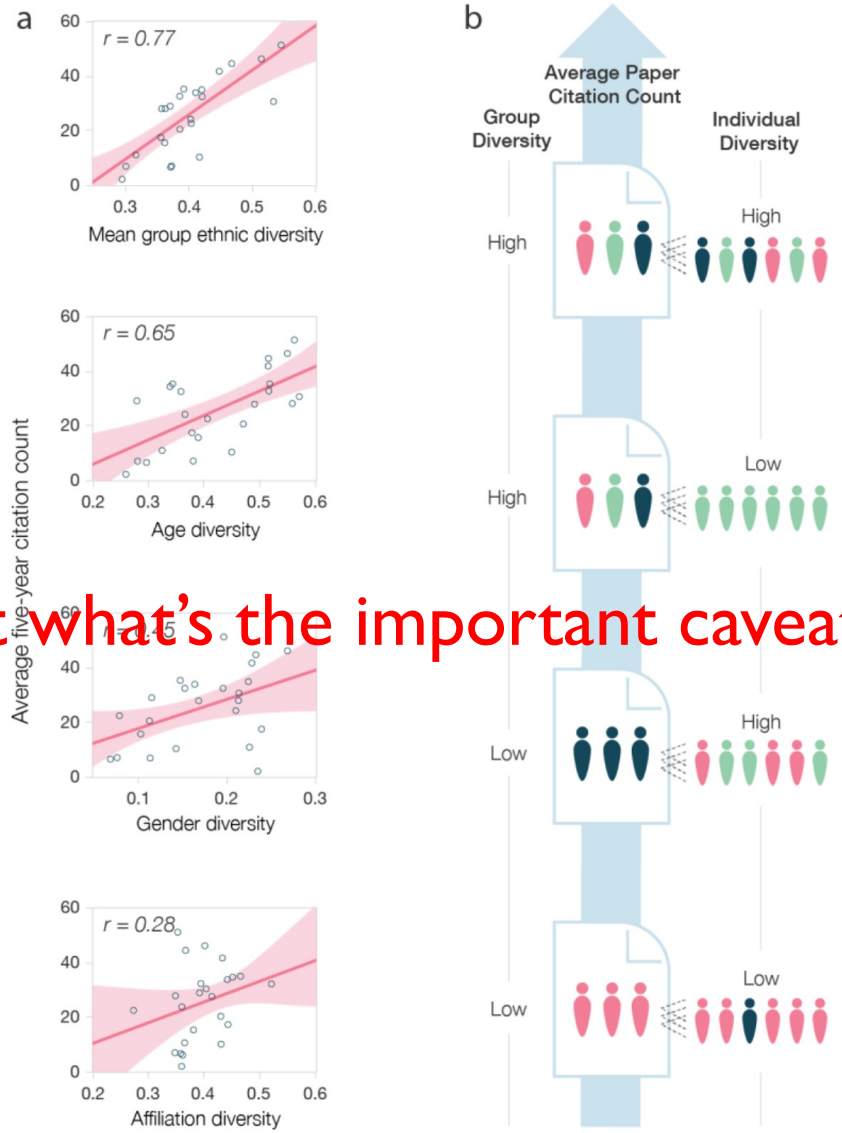
"Relatively weak social ties on LinkedIn proved twice as effective in securing employment as stronger social ties."
– New York Times

"The Science of Science", Part II
The Science of Collaboration
Chapter 11: Team Assembly

Discussion Points

Models…

… and "reality"



But what's the important caveat?!

Figure 2.4.1 **Team diversity and scientific impact.** Panel a: an analysis of more than 1 million papers in 24 academic subfields (circles) shows that ethnic diversity correlates more strongly (*r*) with citation counts than do diversity in age, gender or affiliation. Panel b: Comparing team versus individual diversity reveals that diversity within the list of authors on a paper (team diversity) has a stronger effect on citation count than diversity in a researcher's network of collaborators (individual diversity). After Powell [58].

Simple (newcomer, incumbent) model

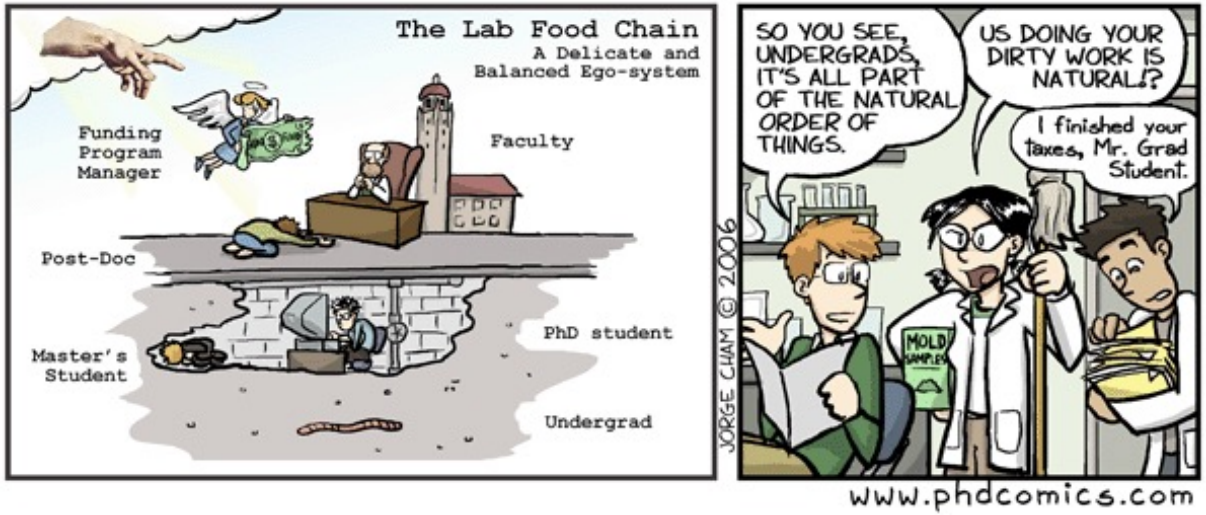
Incumbency parameter p: fraction of incumbents in a team.

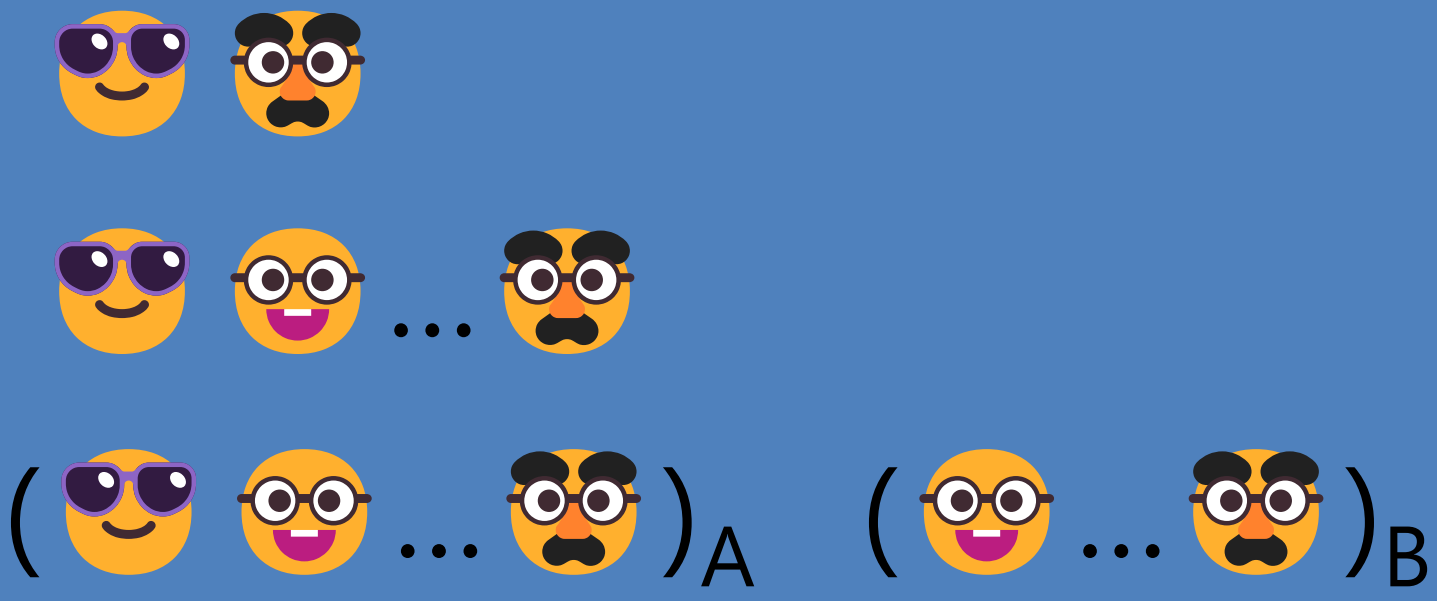
Diversity parameter q: probability incumbents collaborate with prior collaborators. (Higher q means more previous collaborators)

Finding: Journal impact positively correlates with p, negatively correlates with q.

Takeaways?

… but how does it apply to me?


The Lab Food Chain
A Delicate and Balanced Ego-system
Funding Program Manager
Faculty
Post-Doc
PhD student
Master's Student
Undergrad
SO YOU SEE, UNDERGRADS, IT'S ALL PART OF THE NATURAL ORDER OF THINGS.
US DOING YOUR DIRTY WORK IS NATURAL!?
I finished your taxes, Mr. Grad Student.
MOLD SAMPLES
JORGE CHAM © 2006
www.phdcomics.com



(  … )$_A$  $_B$

()$_A$ ( … )$_B$

(  …)$_A$ ( …)$_B$

 $_A$  $_B$

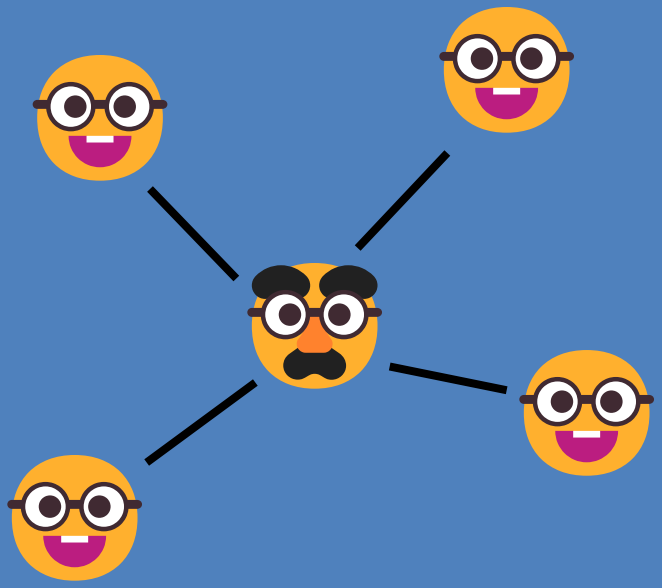
 A VS.  B

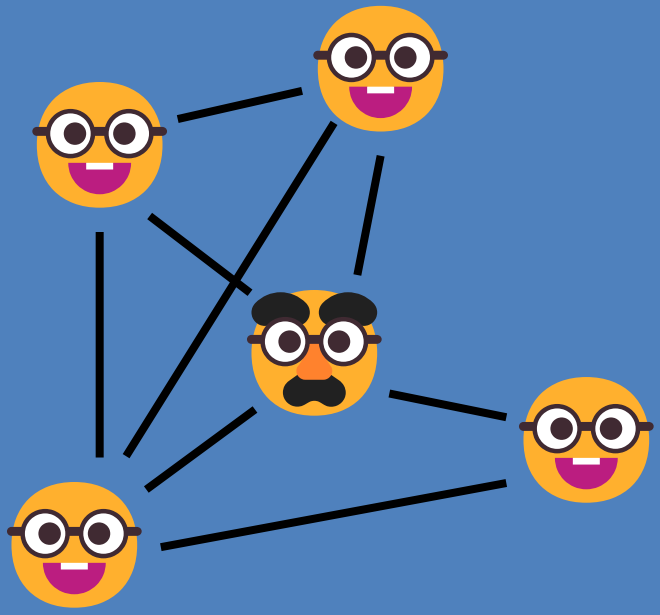
Same organization?

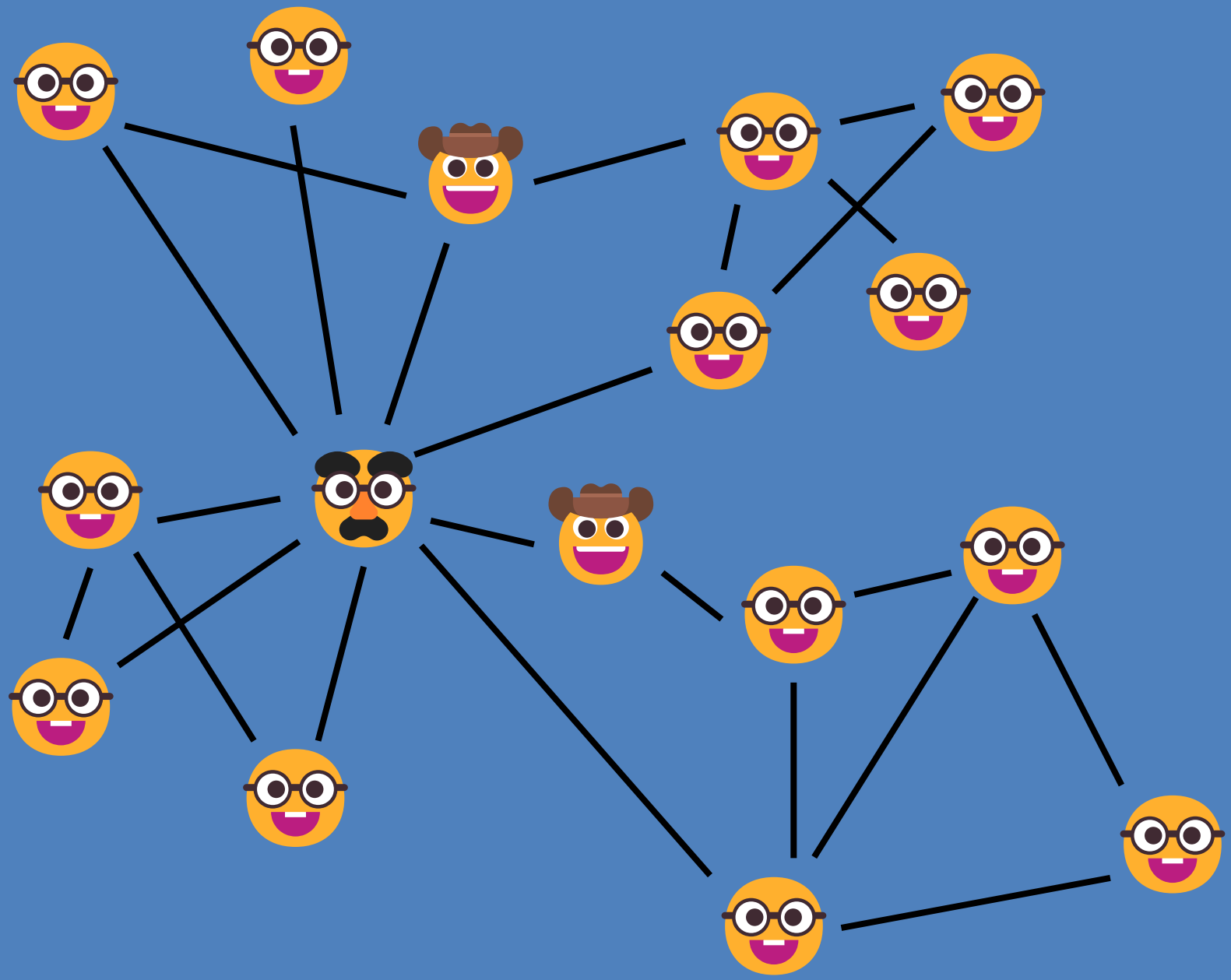
Industry vs. Academia?

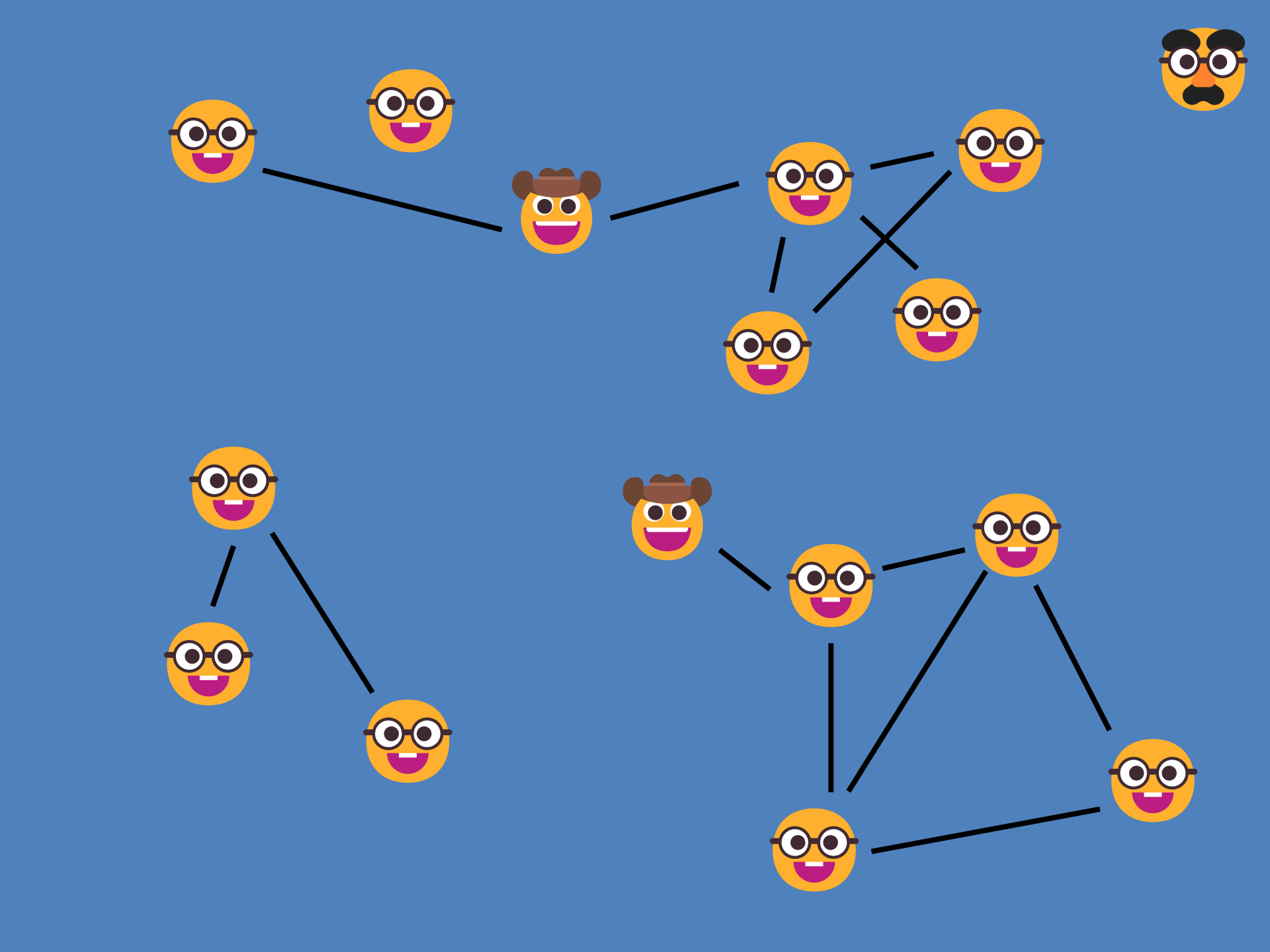
Experience?

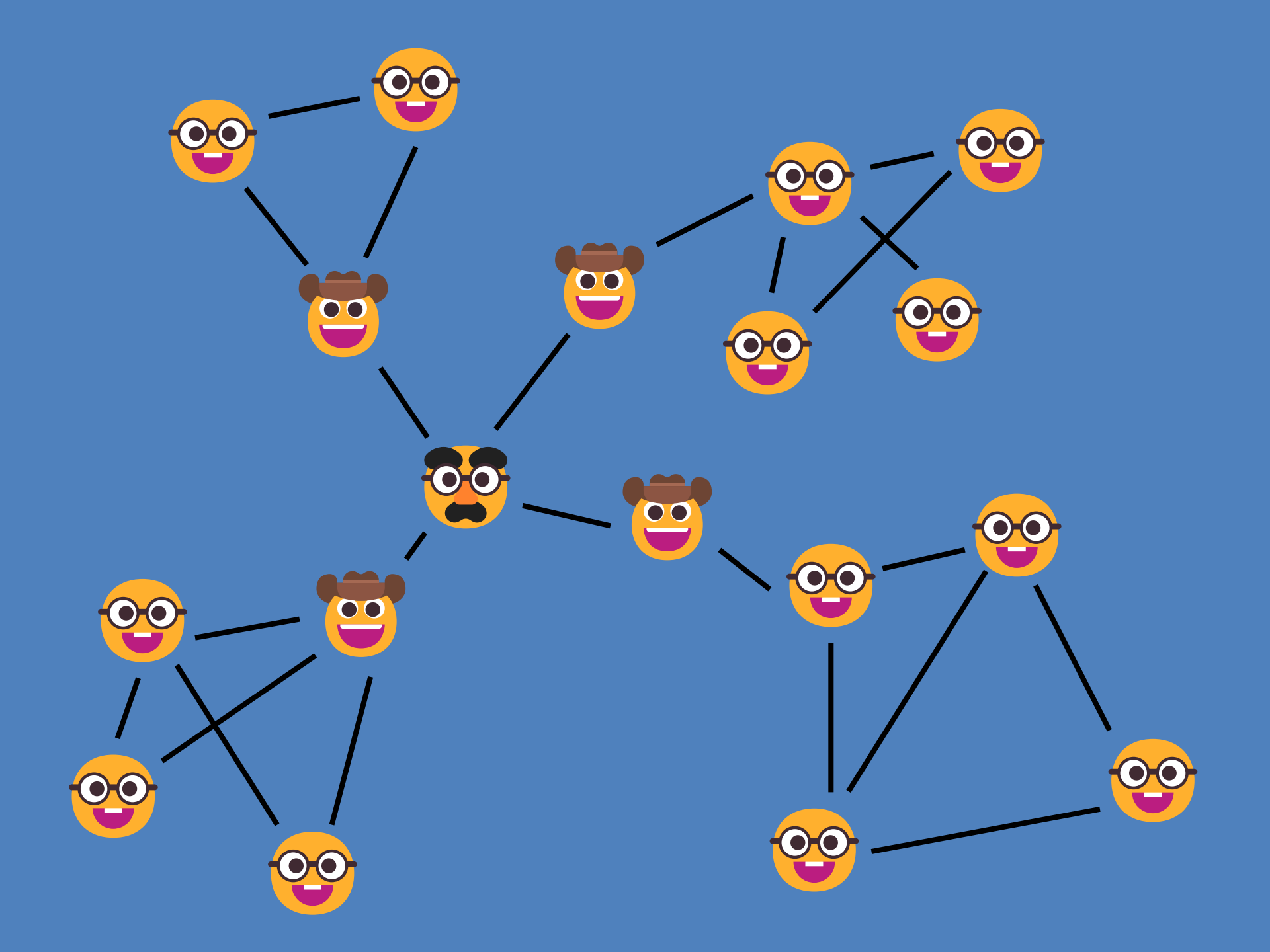
Complementary?











How do teams *really* form?

Source: student vs. advisor

Actual mechanisms?

What should *you* do?

Shockley's Model

$$N \sim p_1\, p_2\, p_3\, p_4\, p_5\, p_6\, p_7\, p_8$$

F_1. Identify a good problem
F_2. Make progress with it
F_3. Recognize a worthwhile result
F_4. Decide when to stop the research and start writing up the results
F_5. Write adequately
F_6. Profit constructively from criticism
F_7. Show determination to submit the paper for publication
F_8. Make changes if required by the journal or the referees

tl;dr – my advice:

Seek homogeneity to share workload

Seek complementarity for everything else

Important: let your advisor know!

"The Science of Science", Part II
The Science of Collaboration
Chapter 12: Small and Large Teams

Discussion Points

Why are teams growing in size?

Top-down vs. Bottom-up

Who determines what research gets done?

"The Science of Science", Part II

The Science of Collaboration

Chapter 13: Scientific Credit
Chapter 14: Credit Allocation

Discussion Points

Normative vs. Positive

Understanding norms

Practical advice

Who gets credit?
Who *should* get credit?
Who *actually* gets credit?

Is it fair?

a VOLUME 76, NUMBER 11 PHYSICAL REVIEW LETTERS 11 MARCH 1996

Generation of Nonclassical Motional States of a Trapped Atom

D. M. Meekhof, C. Monroe, B. E. King, W. M. Itano, and D. J. Wineland

Time and Frequency Division, National Institute of Standards and Technology, Boulder, Colorado 80303-3328

b VOLUME 55, NUMBER 1 PHYSICAL REVIEW LETTERS 1 JULY 1985

Three-Dimensional Viscous Confinement and Cooling of Atoms by Resonance Radiation Pressure

Steven Chu, L. Hollberg, J. E. Bjorkholm, Alex Cable, and A. Ashkin

AT&T Bell Laboratories, Holmdel, New Jersey 07733

c VOLUME 61, NUMBER 21 PHYSICAL REVIEW LETTERS 21 NOVEMBER 1988

Giant Magnetoresistance of (001) Fe/(001) Cr Magnetic Superlattices

M. N. Baibich,[a] J. M. Broto, A. Fert, F. Nguyen Van Dau, and F. Petroff

Laboratoire de Physique des Solides, Université Paris-Sud, F-91405 Orsay, France

P. Eitenne, G. Creuzet, A. Friederich, and J. Chazelas

Laboratoire Central de Recherches, Thomson CSF, B.P. 10, F-91401 Orsay, France

Figure 2.6.1 **Who gets the Nobel?** (A) The last author, David J. Wineland was awarded the 2012 Nobel Prize in Physics for his contribution to quantum computing. (B) Steven Chu, the first author, won the 1997 Nobel in Physics for the paper focusing on the cooling and trapping of atoms with laser light. (C) In 2007, Albert Fert, the middle author of the paper, received the Nobel Prize in Physics for the discovery of the giant magnetoresistance effect (GMR). All three examples are prize-winning papers published in the same journal, *Physical Review Letters*, demonstrating the ambiguity of allocating credit by simply reading the byline of a paper.

NEWS | 22 September 2022

AlphaFold developers win US$3-million Breakthrough Prize

DeepMind's system for predicting the 3D structure of proteins is among five recipients of science's most lucrative awards.

Zeeya Merali

Demis Hassabis (left) and John Jumper (right) from DeepMind developed AlphaFold, an AI that can predict the structure of proteins. Credit: Breakthrough Prize

The researchers behind the AlphaFold artificial-intelligence (AI) system have won one of this year's US$3-million Breakthrough prizes — the most lucrative awards in science. Demis Hassabis and John Jumper, both at DeepMind in London, were recognized for creating the tool that has predicted the 3D structures of almost every known protein on the planet.

Dan Roy
@roydanroy

I was curious how AlphaFold was honored by the breakthrough prize to two individuals. I went back to one of the Nature papers to look at the author contributions. Curious what people see that could justify two people being named.

9:35 PM · Sep 23, 2022 · Twitter Web App

https://twitter.com/roydanroy/status/1573486328502689827

Improved protein structure prediction using potentials from deep learning

Andrew W. Senior ✉, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu & Demis Hassabis

Nature **577**, 706–710 (2020) | Cite this article

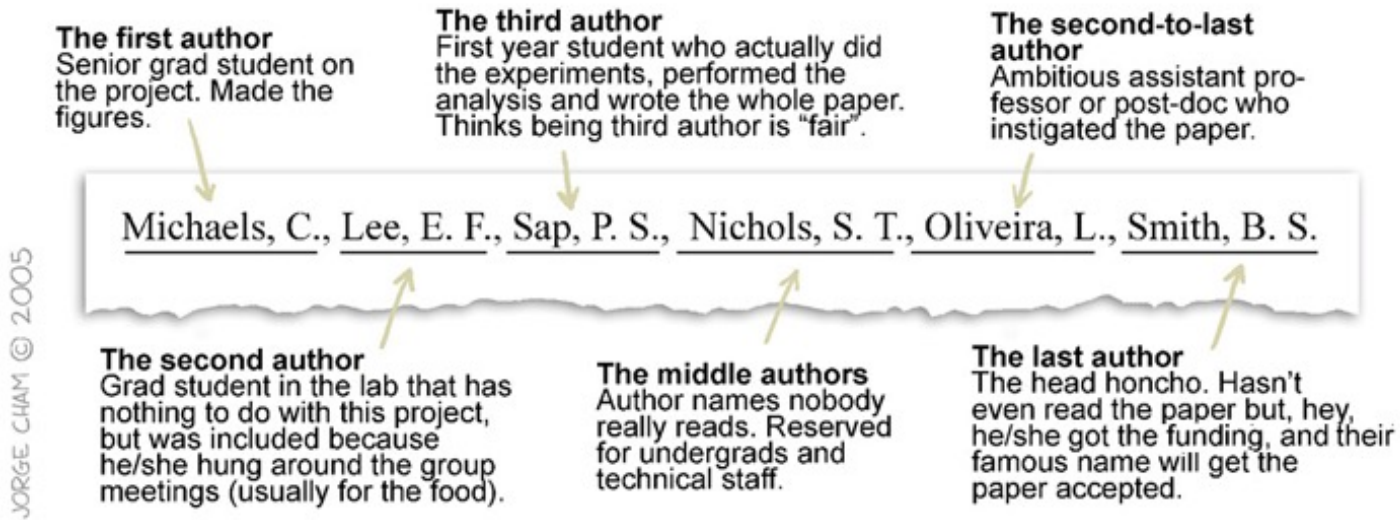
Contributions

R.E., J.J., J.K., L.S., A.W.S., C.Q., T.G., A.Ž., A.B., H.P. and K.S. designed and built the AlphaFold system with advice from D.S., K.K. and D.H. D.T.J. provided advice and guidance on protein structure prediction methodology. S.P. contributed to software engineering. S.C., A.W.R.N., K.K. and D.H. managed the project. J.K., A.W.S., T.G., A.Ž., A.B., R.E., P.K. and J.J. analysed the CASP results for the paper. A.W.S. and J.K. wrote the paper with contributions from J.J., R.E., L.S., T.G., A.B., A.Ž., D.T.J., P.K., K.K. and D.H. A.W.S. led the team.

Corresponding author

Correspondence to Andrew W. Senior.

THE AUTHOR LIST: GIVING CREDIT WHERE CREDIT IS DUE

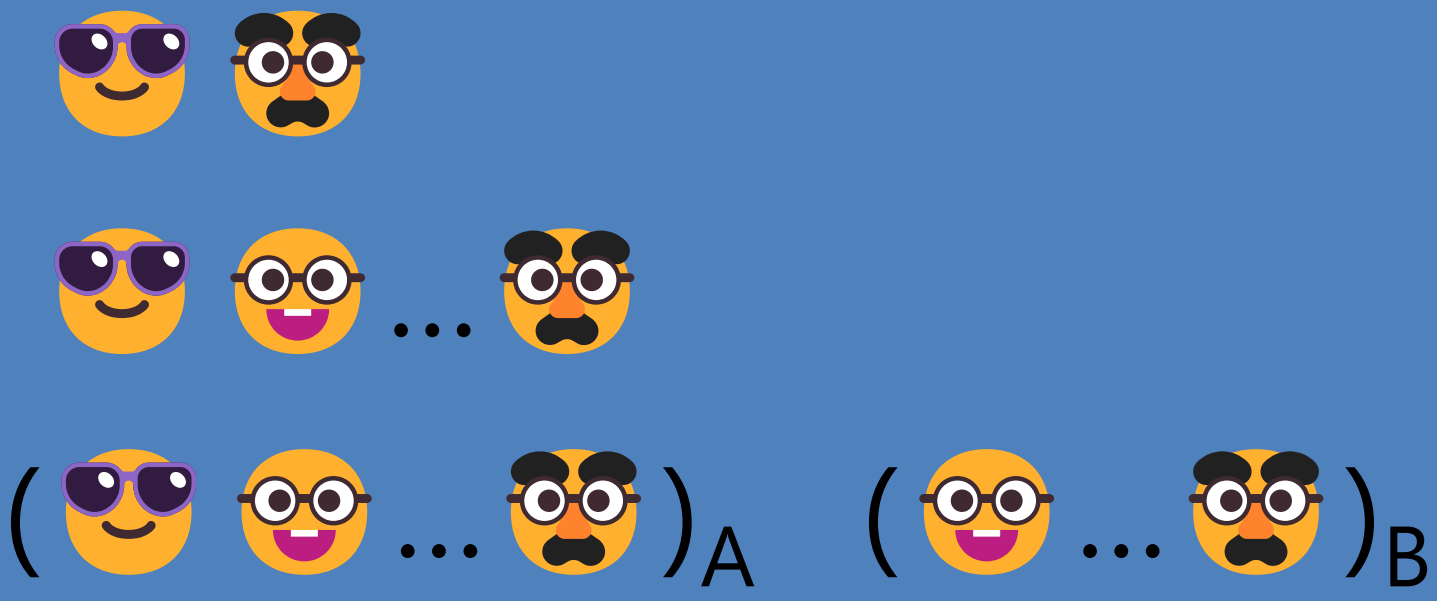


WWW.PHDCOMICS.COM

Ordering of Authors

Field-specific norms (alphabetical vs. contribution-based)
You (as a student) have very limited control

Important first step: understand what the norms are!



(  … )$_A$  $_B$

()$_A$ ( … )$_B$

(  …)$_A$ ( …)$_B$

 $_A$  $_B$

Practical Advice

Determine coauthorship (and ordering) early.
Don't forget to communicate with non-coauthors.
(Personal view) Be generous with the middle author positions.

Who actually gets the credit?

For the most part… your advisor ☹️

Why?

How can you overcome it?

That’s all for this week!