The New York Times

The Opinion Pages
Opinionator
A Gathering of Opinion From Around the Web

ME, MYSELF AND MATH

Friends You Can Count On

By Steven Strogatz September 17, 2012 9:00 pm

Me, Myself and Math, a six-part series by Steven Strogatz, looks at us through the lens of math.

You spend your time tweeting, friending, liking, poking, and in the few minutes left, cultivating friends in the flesh. Yet sadly, despite all your efforts, you probably have fewer friends than most of your friends have. But don't despair — the same is true for almost all of us. Our friends are typically more popular than we are.

Don't believe it? Consider these results from a colossal recent study of Facebook by Johan Ugander, Brian Karrer, Lars Backstrom and Cameron Marlow. (Disclosure: Ugander is a student at Cornell, and I'm on his doctoral committee.) They examined *all* of Facebook's active users, which at the time included 721 million people — about 10 percent of the world's population — with 69 billion friendships among them. First, the researchers looked at how users stacked up against their circle of friends. They found that a user's friend count was less than the average friend count of his or her friends, 93 percent of the time. Next, they measured averages across Facebook as a whole, and found that users had an average of 190 friends, while their friends averaged 635 friends of their own.

Studies of offline social networks show the same trend. It has nothing to do with personalities; it follows from basic arithmetic. For any network where some people have more friends than others, it's a theorem that *the average number of friends of friends is always greater than the average number of friends of individuals*.

This phenomenon has been called the friendship paradox. Its explanation hinges on a numerical pattern — a particular kind of "weighted average" — that comes up in many other situations. Understanding that pattern will help you feel better about some of life's little annoyances.

For example, imagine going to the gym. When you look around, does it seem that just about everybody there is in better shape than you are? Well, you're probably right. But that's inevitable and nothing to feel ashamed of. If you're an average gym member, that's exactly what you should expect to see, because the people sweating and grunting around you are *not* average. They're the types who spend time at the gym, which is why you're seeing them there in the first place. The couch potatoes are snoozing at home where you can't count them. In other words, your sample of the gym's membership is not representative. It's biased toward gym rats.

This is also why people experience airplanes, restaurants, parks and beaches to be more crowded than the averages would suggest. When they're empty, nobody's there to notice.

Weighted averages are the natural measures to use in such cases. An example from the world of education will clarify how they work. Consider a professor who teaches two classes. One is a large introductory course with 90 freshmen in it. The other is an advanced seminar with 10 seniors. What's this professor's average class size?

The university would say 50, because (90 + 10)/2 = 50. The professor would agree. Both of them are implicitly weighting the two classes equally. This is what the usual kind of average does: it assigns half the weight to 90, and half to 10, to arrive at an answer halfway between them. It's not wrong, but in a case like this, it's misleading.

To see why, think about it from a student's point of view. A vast majority of students (90 out of 100) find themselves sitting in a big class of 90. Only 10 experience a class size of 10. Surely that must skew the average from their perspective closer to 90 than 10, and thus above 50.

To calculate this student-weighted average, imagine polling everyone in both classes. When you ask "How big is your class?" 90 students say "90" and 10 say "10." The sum of all their responses equals

And since there are 90 + 10 = 100 students in total, the average class size they experience equals 8,200/100=82, a lot bigger than the average class size of 50 that the university advertises.

The pattern I want you to notice here (please burn it into your neurons; we're going to need it again later) is that 90 and 10 each appear in *two* roles above: as a number being averaged *and* as a weight in front of that number. That's why *two* 90s and *two* 10s appeared in the numerator of the student-weighted average

This same pattern — this dual use of each number — is going to be the key to understanding the friendship paradox.

It's easiest to see how this pattern manifests itself in social networks by looking at a small example in detail. (Nothing I'm about to say depends on the particular structure of the network below; the results are true for any network where some people have more friends than others. But picking a small network makes the math easier to handle.)

In this hypothetical example, Abby, Becca, Chloe and Deb are four middle-school girls. Lines signify reciprocal friendships between them; two girls are connected if they've named each other as friends.

Abby's only friend is Becca, a social butterfly who is friends with everyone.

Chloe and Deb are friends with each other and with Becca. So Abby has 1 friend, Becca has 3, Chloe has 2 and Deb has 2. That adds up to 8 friends in total, and since there are 4 girls, the average friend count is 2 friends per girl.

This average, 2, represents the "average number of friends of individuals" in the statement of the friendship paradox. Remember, the paradox asserts that this number is smaller than the "average number of friends of friends" — but is it? Part of what makes this question so dizzying is its sing-song language. Repeatedly saying, writing, or thinking about "friends of friends" can easily provoke nausea. So to avoid that, I'll define a friend's "score" to be the number of friends she has. Then the question becomes: What's the average score of all the friends in the network?

Imagine each girl calling out the scores of her friends. Meanwhile an accountant waits nearby to compute the average of these scores.

Abby: "Becca has a score of 3."

Becca: "Abby has a score of 1. Chloe has 2. Deb has 2."

Chloe: "Becca has 3. Deb has 2."

Deb: "Becca has 3. Chloe has 2."

These scores add up to 3 + 1 + 2 + 2 + 3 + 2 + 3 + 2, which equals 18. Since 8 scores were called out, the average score is 18 divided by 8, which equals 2.25.

Notice that 2.25 is greater than 2. The friends on average *do* have a higher score than the girls themselves. That's what the friendship paradox said would happen.

The key point is *why* this happens. It's because popular friends like Becca contribute disproportionately to the average, since besides having a high score, they're also named as friends more frequently. Watch how this plays out in the sum that became 18 above: Abby was mentioned once, since she has a score of 1 (there was only 1 friend to call her name) and therefore she contributes a total of 1 x 1 to the sum; Becca was mentioned 3 times because she has a score of 3, so she contributes 3 x 3; Chloe and Deb were each mentioned twice and contribute 2 each time, thus adding 2 x 2 apiece to the sum. Hence the total score of the friends is (1 x 1) + (3 x 3) + (2 x 2) + (2 x 2), and the corresponding average score is

This is a weighted average of the scores 1, 3, 2 and 2, weighted by the scores themselves — the same dual-use pattern as in the class-size problem. You can see that by looking at the numerator above. Each individual's score is multiplied by itself before being summed. In other words, the scores are *squared* before they're added. That squaring operation gives extra weight to the largest numbers (like Becca's 3 in the example above) and thereby tilts the weighted average upward.

So that's intuitively why friends have more friends, on average, than individuals do. The friends' average — a weighted average boosted upward by the big squared terms — always beats the individuals' average, which isn't weighted in this way.

Once this structure has been unearthed, the proof of the rest of the theorem reduces to a matter of algebra (see the notes for the details).

Like many of math's beautiful ideas, the friendship paradox has led to exciting practical applications unforeseen by its discoverers. It recently inspired an early-warning system for detecting outbreaks of infectious diseases.

In a study conducted at Harvard during the H1N1 flu pandemic of 2009, the network scientists Nicholas Christakis and James Fowler monitored the flu status of a large cohort of random undergraduates and (here's the clever part) a subset of friends they named. Remarkably, the friends behaved like sentinels — they got sick about *two weeks earlier* than the random undergraduates, presumably because they were more highly connected within the social network at large, just as one would have expected from the friendship paradox. In other settings, a two-week lead time like this could be very useful to public health officials planning a response to contagion before it strikes the masses.

And that's nothing to sneeze at.

NOTES

1. For a preprint of the study of the Facebook social network, see J. Ugander, B. Karrer, L. Backstrom and C. Marlow, “The anatomy of the Facebook social graph.” Their statistical analysis is much more extensive than my treatment here might suggest, and includes considerations of median versus average friend counts, correlations between a user’s friend count and that of his or her friends, the number of degrees of separation between users, and so on.

2. The friendship paradox was discovered and explained by the sociologist Scott Feld in a paper with a memorable title: S. L. Feld, “Why your friends have more friends than you do,” American Journal of Sociology, Vol. 96, No. 6 (May 1991), pp. 1464–1477.

3. On p. 1470 of Feld’s article, he proves the theorem stated in the friendship paradox by deriving the following identity:

(average number of friends of friends) = (average number of friends of individuals) + (variance in number of friends of individuals) / (average number of friends of individuals).

Since the variance is positive (assuming some people have more friends than others), the theorem follows.

4. For readers who are comfortable with algebraic manipulation, here’s how to derive the identity above. Let x_i denote the number of friends of individual i, for i = 1, 2,…, n, and let

denote the average number of friends of individuals. Then, by definition, the variance of the number of friends of individuals is given by

This can be expanded and simplified to

Dividing both sides by the average of x and rearranging yields

And now we’re done. The left-hand side is the average number of friends of friends. To recognize it as such, notice that it's the weighted average of the type discussed in the main text, where each individual's friend count x_i is weighted by x_i itself; that's why the x's are *squared* in the numerator before they're averaged.

5. Feld, in collaboration with Bernard Grofman, had previously pointed out why students so often find themselves in college classes that are more crowded than average, in S. L. Feld and B. Grofman, "Variation in class size, the class size paradox, and some consequences for students," Research in Higher Education, Vol. 6, No. 3 (1977), pp. 215–222. For an independent and equally insightful take on the same ideas, see D. Hemenway, "Why your classes are larger than 'average,'" Mathematics Magazine, Vol. 55, No. 3 (May 1982), pp. 162–164.

6. The early warning system for detecting outbreaks of flu and other contagious diseases was described in N. A. Christakis and J. H. Fowler, "Social network sensors for early detection of contagious outbreaks," PLoS ONE, Vol. 5, No. 9 (2010): e12948.

7. Besides suggesting a strategy for detecting infection, the friendship paradox suggests a strategy for *combating* it. The idea is to immunize the friends of random nodes, rather than the nodes themselves. See R. Cohen, S. Havlin, and D. ben-Avraham, "Efficient immunization strategies for computer networks and populations," Physical Review Letters, Vol. 91, No. 24 (2003), 247901. Using computer simulations, the authors find that this approach is much more effective than random immunization at halting an epidemic. The technique achieves herd immunity when around 20 to 40 percent of the friend population is immunized, as opposed to the 80 or 90 percent coverage needed when the population at large is immunized. And in a chilling final sentence, they suggest that their strategy might also be relevant to dismantling terrorist networks: "Our findings suggest that an efficient way to disintegrate the network is to focus more on removing individuals whose name is obtained from another member of the network."

Thanks to Margaret Nelson, JoJo Strogatz and Leah Strogatz for preparing the illustration, and Paul Ginsparg, Jon Kleinberg, Andy Ruina, Carole Schiffman and

Johan Ugander for their comments and suggestions.

Limit the Use of My Sensitive Information California Notices