THE FUTURE OF EMPIRICAL FINANCE

Marcos López de Prado [†]

Journal of Portfolio Management, Summer 2015, forthcoming

This version: May 31, 2015

[†] Senior Managing Director, Guggenheim Partners, New York, NY 10017. Research Fellow, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. E-mail: lopezdeprado@lbl.gov. Web: www.QuantResearch.org

I would like to acknowledge useful comments from David H. Bailey (Lawrence Berkeley National Laboratory), José Blanco (Credit Suisse), Jonathan M. Borwein (University of Newcastle), Peter Carr (Morgan Stanley, NYU), Marco Dion (J.P. Morgan), Matthew D. Foreman (University of California, Irvine), David Hand (Winton Capital), Campbell Harvey (Duke University), Attilio Meucci (KKR, NYU), Philip Protter (Columbia University), Riccardo Rebonato (PIMCO, University of Oxford), Luis Viceira (HBS), Ivo Welch (UCLA) and Jim Qiji Zhu (Western Michigan University).

THE FUTURE OF EMPIRICAL FINANCE

ABSTRACT

Empirical finance is in crisis: Our most important discovery tool is historical simulation, and yet, most backtests and time series analyses published in journals are flawed. The problem is well-known to professional organizations of statisticians and mathematicians, who have publicly criticized the misuse of mathematical tools among Finance researchers. In this note I point to three problems and propose four practical solutions. In an attempt to overcome the challenges posed by multiple testing and selection bias, I emphasize the need to move from an individual-centric to a community-driven research paradigm. Low retraction rates can be corrected through technologies that derive "peer p-values". Stronger theoretical foundations and closer ties with financial firms would help prevent false discoveries.

Keywords: Empirical research, false discovery, multiple testing, physics envy.

JEL Classification: G0, G1, G2, G15, G24, E44.
AMS Classification: 91G10, 91G60, 91G70, 62C, 60E.

Empirical finance is in crisis: Our most important discovery tool is historical simulation, and yet, most backtests and time series analyses published in journals are flawed. The problem is well-known to professional organizations of statisticians and mathematicians, who have publicly criticized the misuse of mathematical tools among finance researchers (Bailey et al. [2014, 2015], Bailey and López de Prado [2012, 2014]). The president-elect of the American Finance Association has stated that, for this reason, *"most claimed research findings in financial economics are likely false"* (Harvey and Liu [2014, 2015a, 2015b], Harvey et al. [2015]). These flaws may invalidate a large portion of the work done over the past 70 years. In this article I propose a few general ideas about how empirical finance may be set on a stronger foundation.

PROBLEM #1: MULTIPLE-TESTING

What could possibly rock the empirical finance edifice to the point of collapse? In order to understand where things began to go wrong, we have to go back to the early 20th century. Neyman and Pearson [1933] developed the procedure to test alternative hypothesis, a statistical version of the logical *reductio ad absurdum* argument. First, a researcher states a hypothesis, and then she attempts to reject that claim while accepting a probability of a false discovery α. The goal is to conclude that the alternative hypothesis must be true, with a confidence level $1 - \alpha$.

In 1936, an Italian mathematician named Carlo Bonferroni noted that applying a test of hypothesis multiple times would lead to an increased probability of a false discovery. For example, if we set the confidence level to 95%, the probability of a false discovery (or type-I error) is 5% for the first application, however it gradually increases as we apply the test a second time and so on. Eventually, that probability becomes 100%. This important warning was largely ignored at first on practical grounds: In that pre-silicon chip Era, computational limitations and scarcity of data made multiple testing relatively rare. After these constraints were overcome a few decades later, the problem of multiple testing generated intense preoccupation among the mathematical community. For instance, the American Statistical Society warns against it in its ethical guideline #A.8: *"Selecting the one 'significant' result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading."*

Two celebrated breakthroughs promised to put the problem of multiple testing to rest: First, Hochberg and Tamhane [1987] proposed the Familywise Error Rate method (FWER), which computes the probability of making even one false discovery. This was considered excessively conservative, as having some confidence that no false discoveries occur at all requires very high rejection thresholds. For this reason, Benjamini and Hochberg [1995] proposed the False Discovery Rate (FDR), which computes the probability of generating false discoveries at a rate greater than a user-define tolerance. Both treatments involve raising rejection thresholds as the number of trials increases. FDR is the preferred approach is most non-life threatening applications.

RECOMMENDATION #1: IMPLEMENT PROCEDURES TO DERIVE "PEER P-VALUES"

Leading academic journals in finance routinely publish studies without reporting the number of trials, making it impossible to determine the true probability that the discovery is false. First, all papers could be required to include a section explaining how the authors have controlled for multiple testing. Bailey and López de Prado [2014], Bailey et al. [2014, 2015] and Harvey and Liu [2015a, 2015b] provide several

consistent practical approaches. Second, referees could customarily challenge any claim by requiring the authors to re-test their hypothesis on a few datasets of the referees' choosing. Results would be added to the publication, detailing how the referees attempted to prevent selection bias. Third, journals could set up websites where readers would present their evidence in favor and against the discovery. Authors would also publish there all of their trials, and the software needed to reproduce their results. These forums would allow the discussion of further robustness checks, thus transforming every publication into a coordinated and ongoing collective research project. Fourth, after a period of discussion, qualified participants in the forum would be allowed to openly vote for the retraction of the publication, implicitly deriving a consensus probability of false discovery (a Bayesian-prior p-value, or *"peer p-value"*). The author would retain the decision to withdraw the paper, with the understanding that a high "peer p-value" entails that the community as a whole is recommending a retraction based on the full review of the submitted evidence.

Financial economics is a prolific (Exhibit 1), topic-redundant (Exhibit 2) asocial field (Exhibits 3-4), where most papers go largely ignored (Exhibit 5). A community-driven research paradigm would help reduce the selection bias intrinsic in this individual-centric research approach. The Polymath Projects (polymathprojects.org) offer impressive examples of crowdsourced research.

RECOMMENDATION #2: PROVIDE ETHICAL GUIDELINES

Most modern econometrics textbooks do not mention multiple-testing or its treatment. Econometricians graduate from our schools without knowledge that it is misleading to apply a test of hypothesis more than once, and report the preferred outcome without carrying out a FWER or FDR correction. Researchers perpetuate the convenient myth that hold-out, cross-validation, walk-forward and other schemes control for overfitting and false discoveries. This is obviously not true, as any of these schemes can be repeated as many times as needed for a false discovery to be selected (on average, it takes as few as $1/\alpha$ independent hold-out iterations to produce a false discovery). Out-of-sample testing is meaningful if and only if all trials are reported. A second myth is that simple models are hard to overfit. This is also patently incorrect, as we just need to continue testing until any specification, no matter how simple, will yield to a fluke (a practice sometimes referred to as "p-hacking"). The Econometric Society could officially denounce these unethical practices, much like the American Statistical Society did decades ago.

PROBLEM #2: FINANCE IS AN ADAPTIVE SYSTEM

As soon as we uncover an effect, that effect is likely to disappear due to arbitrage forces (McLean and Pontiff [2015]). Consider what would be the state of physics if Nature worked like a financial system: The publication of Newton's *"Principia"* in 1687 would have unleashed forces that would have cancelled Gravity. Nobody would have been able to verify Newton's equations through experimentation. Some people have compared this situation to "Heisenberg's uncertainty principle", by which the act of observing some quantum variables alters the observed object. In fact, it is worse than that. In finance, the one and only system is permanently altered in the way that most *precisely invalidates* the discovery. Some colleagues have argued that this is not necessarily true of some non-arbitrageable phenomena, like risk premia. However, the discovery of those premia has led to the explosive growth of smart beta and other factor-based products, whereby investors hold passive investment over longer horizons.

Harvesting those risks may eventually dampen those premia. Besides, Goodhart's law exemplifies how non-arbitrageable financial variables (like official statistics) are also impacted by publication.

RECOMMENDATION #3: OVERCOME PHYSICS' ENVY

There are reasons for empirical research to play in finance a smaller role than in physics. First, because one consequence of the above is that signals tend to be barely distinguishable from noise in finance. This makes extremely hard to avoid false discoveries, and easy to discard true discoveries, even after controlling for multiple-testing. Second, because the scientific method was devised to study immutable laws of Nature. It was not devised to study the mutable phenomena of human institutions. We are, after all, the active creators of those financial systems, and we do not need to guess anyone else's grand design. Financial researchers have a much better vantage point with regards to financial institutions than physicists have with regards to Nature.

Empirical research in finance could be mainly concerned with quantifying the parameters of well-thought prior theories, strongly formalized in mathematical terms. Market microstructure research provides a positive example that other financial research areas could try to follow: Most empirical work is conducted to corroborate or quantify a theory that is not founded on experimentation, but on the intrinsic knowledge we have as market designers. Consider how microstructure research approaches the study of prices and volatility. Rather than treating these variables like some sort of alien signal captured in deep space, microstructure researchers incorporate information about the market's structure, participants, goals, matching mechanism, system topology, etc. By doing so, their empirical work is much more constrained by theory and less prone to accept blindly what noise may relay. Market microstructure models are extensively used by high frequency traders, algorithmic traders, broker-dealers, market makers, large execution desks, liquidity analysts, exchanges, regulators, etc. and are among the most successful academic products used by the financial industry.

PROBLEM #3: WHERE ARE THE FINANCIAL LABORATORIES?

Financial academics do not have laboratories where experiments can be repeated under controlled conditions. We cannot repeat the events of the May 6th 2010 Flash Crash in absence of spoofing so as to isolate the cause-effect mechanism. The closest things we have to a laboratory are research-driven financial firms, for three reasons: First, those firms can interact with the system they study. For example, firms deploy an execution algorithm and experiment with alternative configurations. These discoveries are not based on historical simulations, but on trial and error, much like in scientific laboratories. Second, those firms can control for the increased probability of false positives that results from multiple testing. Their research protocols can *legally enforce* the accounting of the results from all trials carried out by employees. Third, true out-of-sample testing occurs only in financial firms. The success of industrial models is objectively tracked and voted by investors. When firms do not make money, they go out of business, hence the firm's strong incentive to backtest carefully and avoid false discoveries.

RECOMMENDATION #4: EMBRACE A MORE BALANCED EDITORIAL MODEL

Historical simulations, whether embodied in econometric models or backtesting, may be more unreliable in academic settings. Retractions in finance are reserved to cases of gross negligence or misconduct, rather than false discoveries in general. The retraction rate of academic papers in finance is extremely low compared to other empirical fields with access to laboratories (Grieneisen and Zhang

[2012]). Compare this to high attrition rate in hedge funds (Getmansky et al. [2004]). Until this state of affairs changes, true discoveries in empirical finance are more likely to come from research-driven firms than academia. Admittedly, only a small portion of (very successful) financial firms follow a research-oriented business model, adhering to strict protocols similar to those implemented in scientific laboratories (see López de Prado [2014] for examples of those protocols).

Most financial journals are edited by academia to serve academia. These journals could promote high quality academic research with strong industry involvement. This would result in a more balanced and cooperative editorial approach, whereby academic theories are tested by financial institutions. The strongest evidence that a discovery is true comes when financial firms make successful use of it.

CONCLUSION

Empirical finance is at risk of becoming a pathological science, a collection of "cold fusion" claims. In an attempt to overcome the challenges posed by multiple testing and selection bias, I have emphasized the need to move from an individual-centric to a community-driven research paradigm. Artificially low retraction rates can be corrected through technologies that derive "peer p-values". Stronger theoretical foundations and closer ties with financial firms would help prevent false discoveries.

Ten years ago, medical research was in a juncture similar to the one faced by finance today. In their case, the existence of laboratories made it embarrassingly clear that new research protocols were needed, and government agencies threatened action. Financial firms are not in the business of debunking false discoveries published in academic journals, and so the status quo has been kept for far too long. However, three disruptive technologies are about to make these problems even worse: First, Big Data sets are proliferating, with financial markets, business and social media generating many terabytes of data per day. Second, machine-learning algorithms allow for a wide variety of parametric and non-parametric specifications to be fitted and tested, regardless of theoretical consideration. Third, high-performance computing facilities allow us to run billions of those tests on these datasets and specifications every day. The consequence will be an ever greater proliferation of all sorts of false claims, statistical flukes without theoretical support, each of them with infinitesimal p-values. Time is of the essence.

Prof. Campbell Harvey has been a thought leader on these issues. As his tenure as president of the American Finance Association is about to begin, there is hope that this crisis will be finally confronted. Ultimately it is up to us what kind of finance we want to devote our careers to.

REFERENCES

- Bailey, D., J. Borwein, M. López de Prado and J. Zhu (2014): "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-Of-Sample Performance", **Notices of the American Mathematical Society**, 61(5): 458-471. Available at http://ssrn.com/abstract=2308659
- Bailey, D., J. Borwein, M. López de Prado and J. Zhu (2015): "The Probability of Backtest Overfitting", **Journal of Computational Finance**, forthcoming. Available at http://ssrn.com/abstract=2326253

- Bailey, D., M. López de Prado (2012): "The Sharpe Ratio Efficient Frontier", **Journal of Risk**, 15(2): 3-44. Available at http://ssrn.com/abstract=1821643
- Bailey, D., M. López de Prado (2014): "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality", **Journal of Portfolio Management**, 40(5): 94-107. Available at http://ssrn.com/abstract=2460551
- Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing", **Journal of the Royal Statistical Society**, Series B, 57(1): 289–300.
- Getmansky, M., A. Lo and S. Mei (2004): "Sifting through the wreckage: Lessons from recent hedge fund liquidations", **Journal of Investment Management**, 2(4): 6–38
- Grieneisen, M and M. Zhang (2012): "A Comprehensive Survey of Retracted Articles from the Scholarly Literature", **PLOS One**. Available at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0044118
- Harvey, C. and Y. Liu (2014): "Evaluating Trading Strategies", **Journal of Portfolio Management**, 40(5): 108-118. http://ssrn.com/abstract=2474755
- Harvey, C. and Y. Liu (2015a): "Lucky Factors", working paper. Available at http://ssrn.com/abstract=2528780
- Harvey, C. and Y. Liu (2015b): "Backtesting", working paper. Available at http://ssrn.com/abstract=2345489
- Harvey, C., Y. Liu and H. Zhu (2015): "… and the Cross-Section of Expected Returns", working paper. Available at http://ssrn.com/abstract=2249314
- Hochberg, Y. and A.C. Tamhane (1987): **Multiple Comparison Procedures**. New York: Wiley.
- López de Prado, M. (2014): "Quantitative Meta-Strategies", **Practical Applications**, 2(3): 1-3.
- McLean, R. and J. Pontiff (2015): "Does Academic Research Destroy Return Predictability?", **Journal of Finance**, forthcoming. Available at http://ssrn.com/abstract=2156623
- Neyman, J. and E. Pearson (1933): "On the Problem of the Most Efficient Tests of Statistical Hypotheses", **Philosophical Transactions of the Royal Society A**, 231(694–706): 289–337.

EXHIBITS

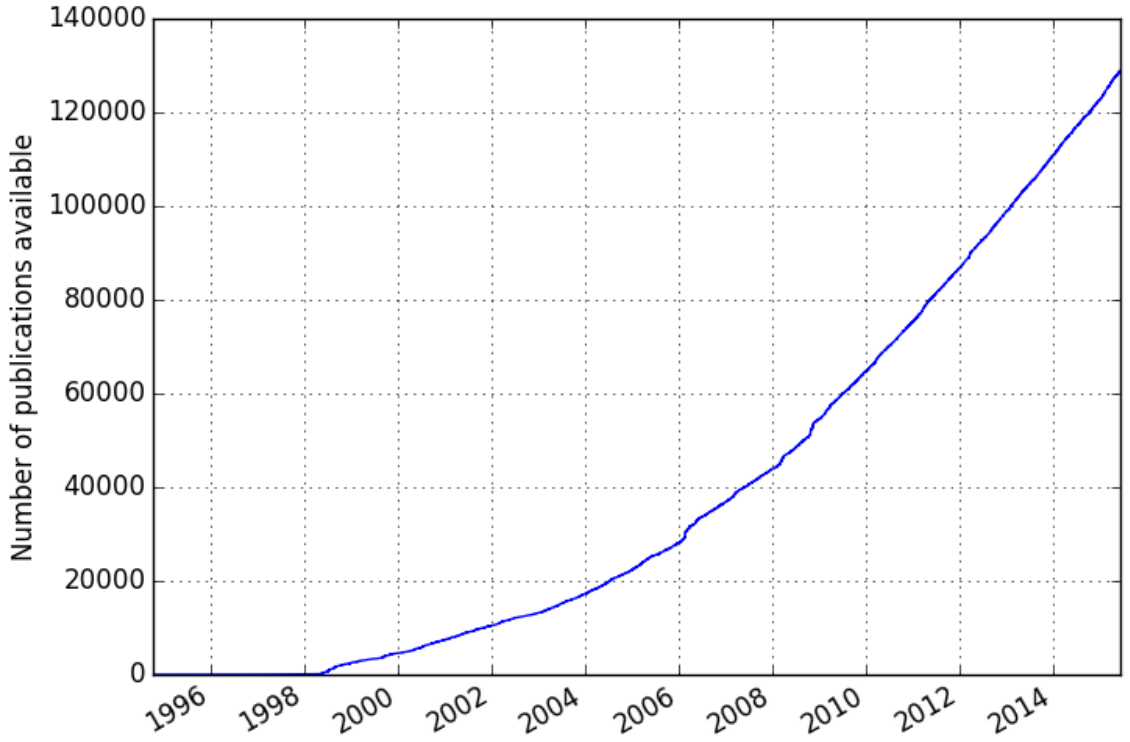

Exhibit 1 – Number of publications in financial economics

Using a web-scraping program, I have extracted information from all of the 128,897 research papers (as of June 4th 2015) published in SSRN's Financial Economics Network (FEN). There has been an explosive growth of papers in recent years.

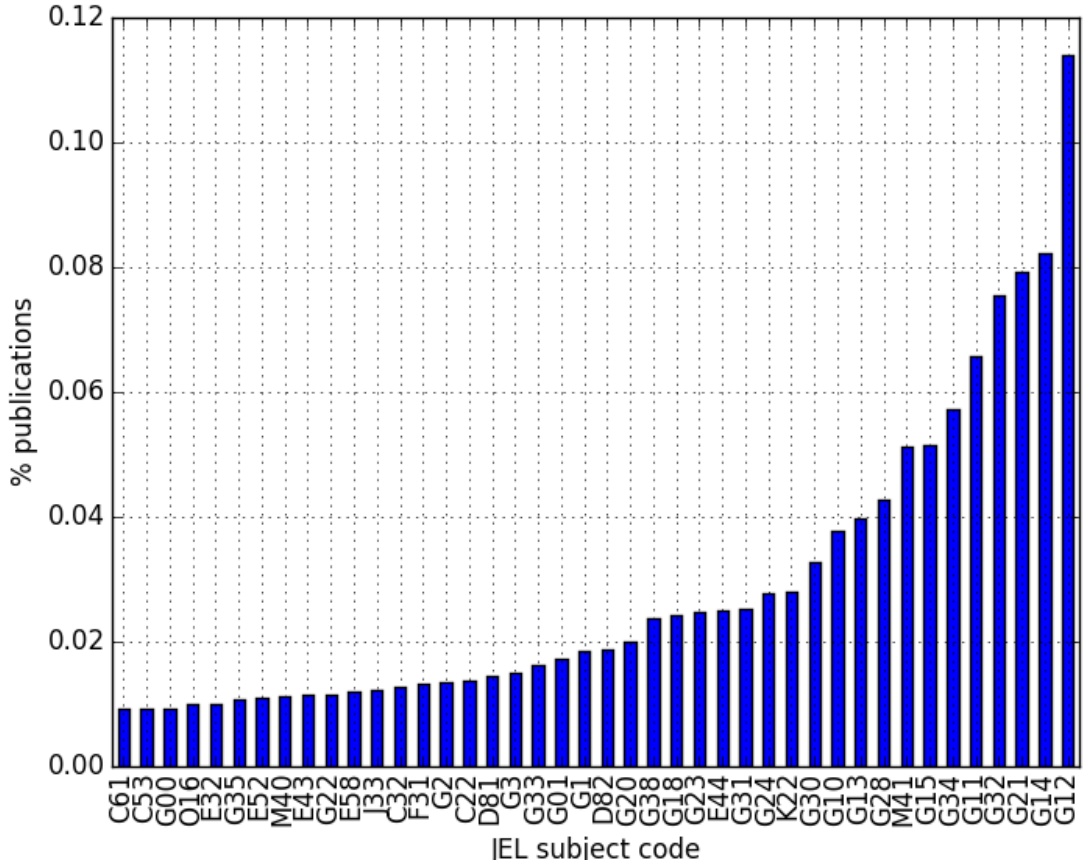

Exhibit 2 – Percentage of publications per subject

Authors assign and average of 2.06 JEL subject codes to each paper, so each paper is relatively focused on a particular subject. The most popular subject by far is asset pricing (G12), followed by market efficiency (G14), banking (G21) and risk management (G32). These subjects make frequent use of time series analysis and backtesting, with the problems discussed in this article.

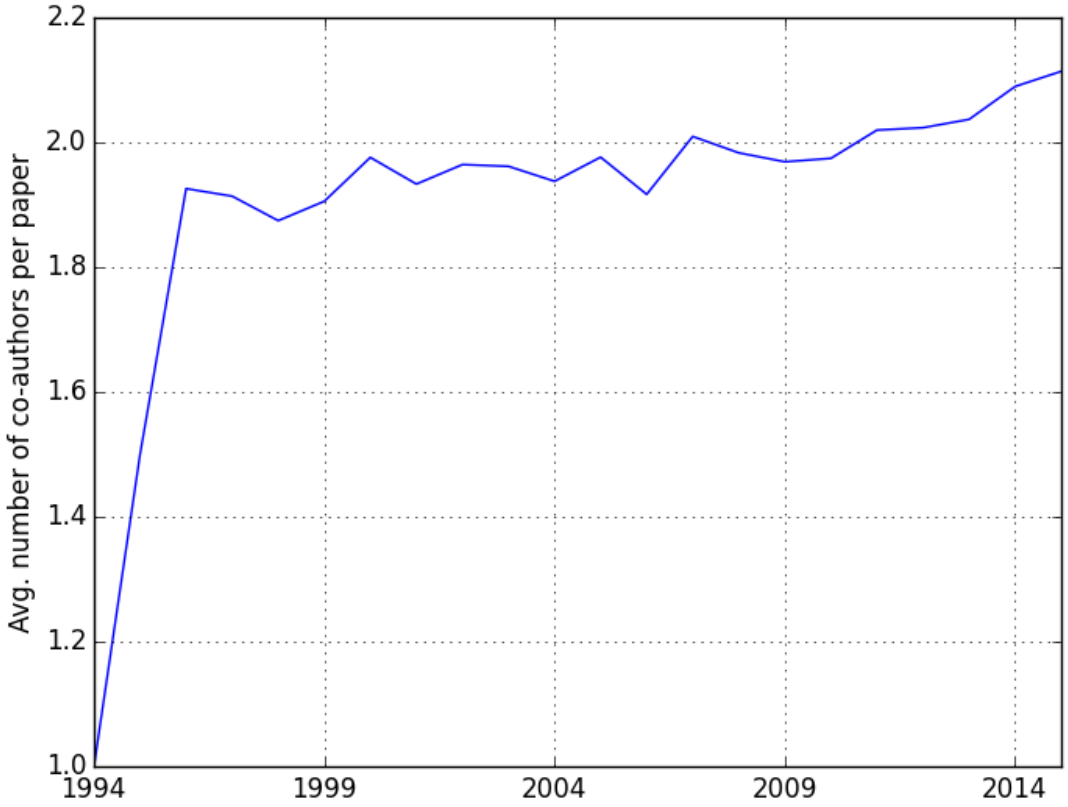

Exhibit 3 – Average number of co-authors per paper over time

Co-authorship across the 128,897 papers is very low compared to other research fields. For example, in medicine the average number of co-authors per paper is 5.5, and only 3% of papers are single-authored.[1] In contrast, papers in financial economics have only 2.00 co-authors on average, and 35.85% of papers are single-authored, a high level compared to experimental disciplines.[2]

[1] http://www.nlm.nih.gov/bsd/authors1.html
[2] http://chronicle.com/article/A-Science-Leaves-the-Solo/142903/

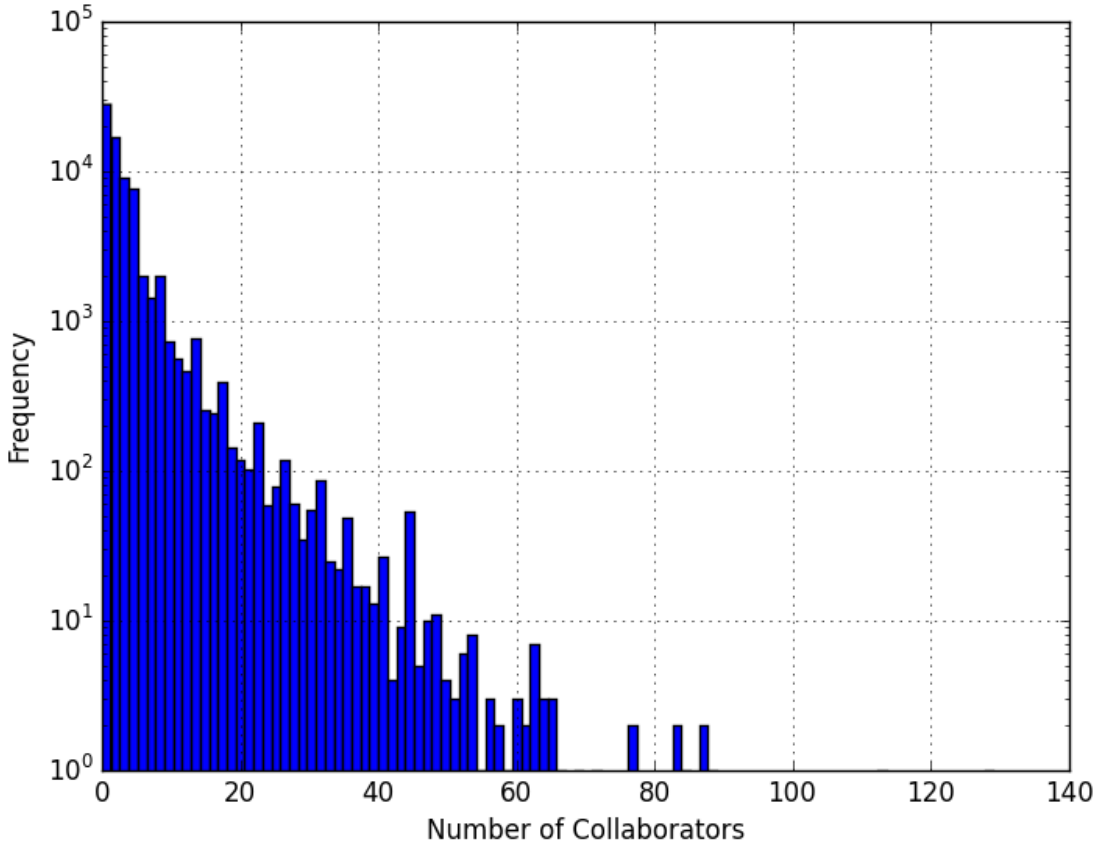

Exhibit 4 – A surprisingly asocial social science

The degree of collaboration across the 72,070 authors is also extremely low. About 13.91% of authors are lone wolves (zero collaborators). Only 0.64% of authors have 30 or more collaborators across all FEN papers. The most "social" financial economist has 129 collaborators. In comparison, mathematician Paul Erdős had 511 collaborators.

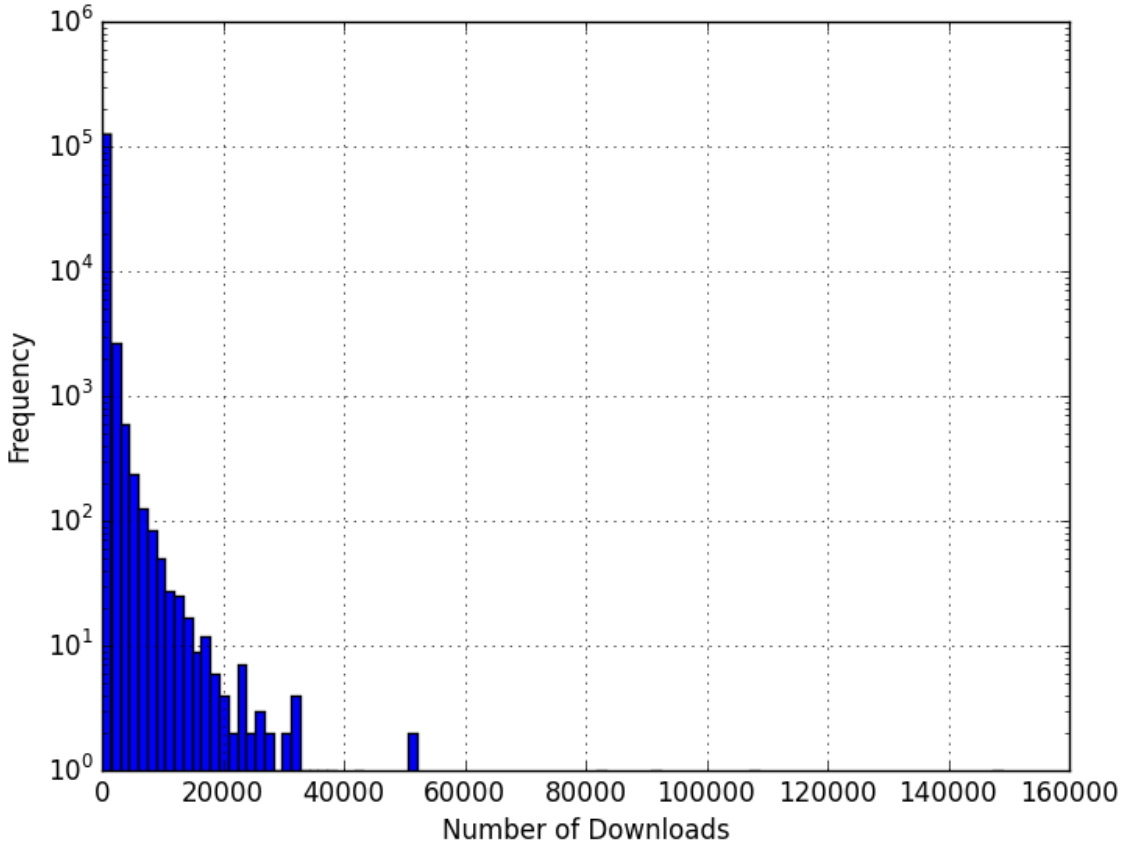

Exhibit 5 – A small number of highly influential papers

Despite the proliferation of publications, the number of downloads is highly concentrated. Downloads per quantile are: 12 (Q1), 72 (Q2) and 242 (Q3). The median number of downloads across the bottom 90% papers is only 56, while the top 10% papers receive a median of 1,053 downloads. The most popular paper has received 148,912 downloads. Only 144 papers (0.11%) have received more than 10,000 downloads.