COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK
ECONOMICS

November 22, 2015

Dear Recruitment Committee,

I am writing to recommend Elliott Ash for academic jobs in economics and law. I have known Elliott since his third year in the Ph.D. program, when he approached me about being interested in using text-analysis tools to analyze the Congressional Record. Since then we have jointly begun a large number of text-analysis projects, effectively operating a lab with 12 RAs, 2 programmers, and a number of projects jointly with Bentley Macleod. I have worked with him quite closely on his job market paper, which I describe below.

Elliott is the best applied microeconomics Ph.D. in the Columbia cohort for this year, and one of the 2 best students I have advised closely since coming to Columbia. Firstly, on raw energy and focus, Elliott is incredibly impressive. Basically he's a dynamo, executing projects at a speed that leaves me shaking my head. We will talk in the morning about an idea, and by the next day he'll have running code, and preliminary results in a week. In terms of programming ability and familiarity with the latest in data manipulation tools, I would put him at the top of the Ph.D. cohort at Columbia. Combined with his legal training and deep substantive interests in political economy and applied microeconomics, I think Elliott stands out from the pack in terms of a distinctive research agenda, as evidenced by his job market paper.

Elliott's job market paper is just one of 8 completed papers he has done since arriving at Columbia, and it is truly creative, bringing together ideas from political economy, law and economics, public finance, machine learning, and applied microeconomics. Elliott's job market paper operationalizes the idea of the "effective tax code" as an instrument for redistribution. Elliott uses natural language processing tools to represent the language of state tax codes, going back to 1963, as a long vector of syntactically correct phrases. This yields an enormous dataset of phrases constituting the tax law, where the frequency of each phrase in each state-year is measured from machine-readable scans of the new tax statutes from a given state and year. Elliott then links this to tax revenue data at the state-year level for sales tax, income tax, and corporate taxes, and measures the tax revenue relative to various measures of the tax-base, conditional on the tax rate(s). This is a measure of the fiscal capacity of a state, measuring how much the state collects relative to the potential tax base. Elliott innovates in a few directions in the details of how this is done, for example restricting attention to phrases that have meaningful linguistic structure (e.g. instead of all 3-word phrases, he takes sequences of adjectives or adverbs followed by a noun or verb.).

The straightforward thing to do would be to just run LASSO or another high-dimensional regression technique to extract which phrases are correlated with fiscal capacity. Even this needs to be modified to account for unobserved heterogeneity. Elliott's basic

1022 International Affairs Building 3308 420 West 118th Street New York, New York 10027 212-854-3680 Fax 212-854-8059

regression uses first-differences with year X tax base FE, thus controlling for state X tax base fixed effects while preserving sparsity of the phrase counts (a point he noticed that seems to be generally useful). He also controls for a wide variety of business-cycle covariates that could alter the size of the tax base.

But of course, endogeneity concerns remain, and so what Elliott does is leverage a commonly used Bartik identification strategy: he uses the lagged average phrase frequency in all the other states that share a judicial district with a given state. The identification is thus obtained by diffusion of shared judicial precedents and common legal culture that is captured in similarities in tax language within judicial distributions. This, together with the year and state X tax base fixed effects, provides a large (over 300K) set of instruments, one for each phrase. Elliott then brings in a high-dimensional 2SLS estimator (essentially an l-1 penalized IV regression developed in statistics) to estimate this. As a robustness check on the required sparsity assumption, he also uses the principal components of the set of frequencies as instruments, building on literature from factor models (Bai and Ng 2002).

With this setup, Elliott is able to estimate the causal effect of tax law language on fiscal capacity. The mechanism makes sense: beyond exemptions, tax accountants and tax lawyers have interpretations of tax law that can severely affect a state's ability to collect revenue. An economist naturally thinks of the tax code as a function mapping assets and sources of income to tax bills. But lawyers worry a lot about the precise wording of tax law, and there is an extensive tax jurisprudence and case law debating and redefining key terminology in the tax code. For example, the phrase "expenses reasonably related to business" could be interpreted by tax authorities very differently, and a skilled tax lawyer could find ways to classify a broader set of expenditures as deductible, reducing the tax revenue collected.

With these phrase-specific IV coefficients, Elliott then constructs the "effective tax code" as the predicted tax revenue from the frequencies of the phrases alone. This is akin to the measure of "slant" in Genktzow and Shapiro (Ecma 2010), although using the Bartik variation rather than possibly endogenous variation (i.e. there could be many other things correlated with Republican and Democrat that the phrases are correlated with), essentially forming a summary measure of the revenue impact of the textual features of the tax code.

Elliott then uses this predicted tax revenue as the outcome in a differences-in-differences design to estimate the effect that Democrat control of state government has on the "effective tax code". The puzzle in this literature is that it is difficult to find an effect of partisan control of state government on tax rates, and while there is some positive effect of Democrat control on tax revenue, the mechanism is unclear. Elliott shows that partisan differences in state power generate subtle differences in wording of the tax code, and these generate increases or decreases in revenue.

The pattern is also consistent with economic incidence. Democrats change the effective tax code to generate more revenue for income and (somewhat) corporate taxes, and less revenue from sales taxes, reflecting the greater progressivity of the former.

Elliot's results are also robust and well-identified, besides ruling out pre-trends and leads and lags, he approximates a regression discontinuity design by controlling for polynomials in seat shares above and below 50% thresholds in both the legislature and the senate seat shares. I think this portion of the job market paper shows that parties do indeed matter for redistribution, but the instruments that they use are not (politically salient) tax rates, but instead subtle differences in wording.

To explore how subtle the differences in wording are, Elliott looks at a "phrase-level" regression, where he regression how much a phrases increases revenue on how much Democrats increase the frequency of a phrase, controlling for a wide variety of phrase-specific characteristics, in particular the contexts in which the phase is used. If Democrats and Republicans just differed in the broad use of language, then controlling for the topics or contexts would eliminate the correlation between Democrat use and additional revenue. However, Elliott finds that this correlation is robust to a large set of topic fixed effects, suggesting that the partisan differences in tax code wording are indeed quite granular. This component of the paper also tells us what to look for in future research: expert scholars can code the tax policies Elliott's inductive approach suggests as important, and look for specific identifying variation in those components.

These results are also substantively important. Elliott decomposes the effect of Democrats on revenue from each source into effects on rate, tax code, and "other policies", and finds that the tax code explains much more of the effect of Democrats on revenue than the rate does. While this section needs more fleshing out, it is interesting that tax rates, the primary focus of public finance, are not what seems to matter for redistribution. Instead it is the subtler definitions of the tax base and other tax regulations that seem to do more work, and this should be a larger focus for economists.

Another exciting thing this paper leads to is the possibility of data driven tax law design. Armed with his estimates of which dimensions of language affect tax collections, Elliott can use new tools in computational linguistics (I describe some of them below) to effectively re-write the tax code, substituting phrases that are linguistically and legally close, but differ substantially in the resulting tax take. Elliott generates model legislation that has a number of phrase substitutions, for example replacing ambiguous uses of the term "lien" with the more specific "lien upon property", increases tax collections by 50 million dollars a year, and replace similarly ambiguous uses of "amount" with "equal total" increases tax collections by around $17 million. While obviously preliminary, I think this is an extremely exciting direction for law and economics: taking language seriously, and quantitatively, and using tools from computer science to help design better policy, not just abstractly but with a concrete understanding of the concerns of tax administration.

As you can see this is quite out-of-the-box and innovative. It is quite high risk, as Elliott is developing many of the tools himself and there is not a canned methodology to follow. But his other advisors and I decided (rather late) that this was what he was passionate about and so he should go forth with it as a job market paper, rather than other projects that, while solidly identified applied microeconomics, public finance or political economy, were not as creative and distinctive. So while the paper is a bit behind on polish right now, in its completed form I expect it to have a good shot at a top journal like Econometrica.

Elliott is part of a generation of applied microeconomists that have internalized the tools of data scientists, with tools for prediction and dimension reduction (often borrowed from time-series econometrics) playing a prominent role in making large unstructured data make sense. In Elliott's case it is natural language processing applied to policy relevant documents such as contracts, court opinions, legislative statutes, the Congressional Record, and of course tax law. Elliott is on top of the latest in computational language processing, including cutting-edge tools such as Word2Vec (a recent method developed by researchers at Google) and sparse topic models. To get into more detail, Word2Vec is a tool that represents a document as a matrix of words and associated "contexts", where each context is a list of words that appear immediately adjacent to the word. This matrix is then factored into a much lower dimensional matrix with latent columns C, so that each word is mapped into a vector of length C. These vectors do an amazing job of capturing meaning of words, and have an uncanny ability to capture analogies. For example: the vector for "King" minus the vector for "Man" plus the vector for "Woman" gives the vector for "Queen". Elliott does the same for terms like "corporate income tax", finding the phrases that are close to it in context space. Elliott is applying these tools to explore a whole new terrain in economics: what role does legal language itself have on economic outcomes.

Elliott is not just applying tools developed by computer scientists. He is thinking carefully about the econometric properties of these estimators, and paying attention to both causality and inference, which tend to be neglected in the NLP applications. For example, an important step is "hard-thresholding" features to obtain a smaller set of phrases (only on the other of a few hundred thousand) that are good predictors of an outcome variable. Gentzkow and Shapiro (2010) do this using a chi-squared statistic calculated for each phrases, keeping only the ones that are statistically significant. As I mentioned above, what Elliott does instead in his job market paper is use the IV coefficients from a regression that uses Bartik-style instruments for phrase frequency, thus restricting the set of predictors to those that identified. This is the first application I know of applying dimension reduction to causal variation with large numbers of both endogenous variables and instruments. In addition, Elliott is also interested in structural econometric models that could give rise to some of these estimators (e.g. some topic models use multinomial logit distributions which naturally have random utility microfoundations.)

Elliott has a large number of other papers, including another potentially promising alternative job market paper that uses exogenous variation in property tax assessment schedules to estimate the effect of local property taxes on economic activity. Interesting, different states generate variation in whether the changes in local tax revenue affect local spending. Using this, Elliott finds that the property tax increases population and new business formation when the proceeds are spent by local government, but has negative effects when the increase in tax revenue goes to a higher level of government and is not spent locally. Elliott relates these empirical elasticities to theoretical results on the optimal property tax, including the "Henry George theorem" by Arnott and Stiglitz (1978).

Elliott has also been a de facto PI on an NSF grant Bentley Macleod and I have on digitizing and analyzing the universe of 20th century labor union contracts. We learned that Cornell has these contracts sitting in boxes in the library, and so we thought a useful public good would be to digitize them all, and then use our NLP toolkit to help make sense of this vast trove of unstructured data. We already have some preliminary results: for example we find that the language from NLRB court opinions diffuse into the language of contracts, and labor lawyers resolve uncertainties and ambiguities by drawing on the court language. We see this dataset as an extremely useful lens on the theory of incomplete contracts, where we use the law and the language of the contract to determine what is left unwritten and how that can generate losses in productivity (measured by stock prices) or even labor conflict (e.g. strikes).

I recommend that all schools look at Elliott, including top 10 economics departments and the very best law schools looking in law and economics. As can be seen from his CV, he is in high demand as a collaborator by faculty at Columbia. Indeed this is perhaps his biggest flaw: he is only too willing to deploy his substantial work capacity to help others around him, from undergraduates to other graduate students to faculty, and this sometimes distracts him from his own projects. He has an extremely productive and promising career ahead of him, and will bring substantial externalities, both social and intellectual, to any department lucky enough to get him.

Please feel free to contact me with any questions by phone at (212)-854-0027 or via email at sn2430@columbia.edu.

Best,

Suresh Naidu

1022 International Affairs Building 3308 420 West 118th Street New York, New York 10027 212-854-3680 Fax 212-854-8059