Prospectus

Elena Khusainova

1 May 2017

1 Introduction

During my second year I took two courses that influenced me most: in the fall semester I took Dan Spielman's course on spectral graph theory and in the spring semester I took a course taught by Van Vu on random matrices. Since then I've been working on several problem involving networks, graphs and random matrices.

First, I started working with Harry on the problem of spam detection based on Kyng et al. (2015) and Zhou et al. (2007). Almost at the same time I was exploring general network theory (Davis-Kahan theorem, random matrix theory, spectral theory) with David to strengthen my background. One of the papers I was particularly interested in was Vu (2007) on behavior of spectral norm of a random matrix.

I made some progress with spam but found myself more attracted to the underlying theory. Thus I started exploring the related literature more intensively. Eventually I concluded that the state of the art is due to Péché and Soshnikov (2007), which is a part of a sequence of several papers written by Soshnikov and his coathors. These papers developed an elaborate technique to address various problems of the field. Unfortunately Péché and Soshnikov (2007) relied heavily on the first paper, by Sinai and Soshnikov (1998a). Even more unfortunately, David and I found in that first paper what appears to be a significant gap in the argument (see section 2). Reluctantly, we decided to put on hold our study of Soshnikov's approach, at least for a moment, although I still wanted to work on networks.

Fortunately Yihong then suggested the problem of RDS (respondent driven sampling), a method that appears to be much used in the social science literature despite the lack of theory. Since February I have been exploring with simulations the estimators and algorithms being used and it appears to be promising (see section 3).

2 From Soshnikov to Vu and back again

The spectral norm $\|A\|_2$ of an $n \times n$ symmetric matrix A is defined as

$$\|A\|_2 := \sup_{\|u\|_2=1} \|Au\|_2.$$

If the eigenvalues of A are labeled in decreasing order of magnitude, $|\lambda_1(A)| \geq |\lambda_2(A)| \geq ... \geq |\lambda_n(A)|$, then $\|A\|_2$ equals $|\lambda_1(A)|$.

Vu (2007) considered the case where A is random and symmetric with (i,j)th element ξ_{ij} satisfying:

$$\mathbb{E}\xi_{ij} = 0,\ var\ \xi_{ij} = \sigma^2,\ |\xi_{ij}| < K,$$

for some constant K. He proved that there is a positive constant $c = c(\sigma, K)$ such that

$$\|A\|_2 \leq 2\sigma\sqrt{n} + cn^{1/4} \ln n,$$

with high probability.

The $2\sigma\sqrt{n}$ factor is sharp. See, for example, Tao (2012, section 2).

To upper bound the spectral norm Vu used the Wigner method, which involves calculating the trace of powers of the matrix for even k:

$$\lambda_1(A)^k = \lambda_1(A^k) \leq \sum_{i=1}^{n} \lambda_i(A^k) = trace(A^k) = \sum_{i_0=1}^{n} \sum_{i_1=1}^{n} ... \sum_{i_{k-1}=1}^{n} \xi_{i_0 i_1} \xi_{i_1 i_2} ... \xi_{i_{k-1} i_0},$$

If we take the expectation of both sides then:

$$\mathbb{E}\lambda_1(A)^k \leq \sum_{i_0=1}^{n} \sum_{i_1=1}^{n} ... \sum_{i_{k-1}=1}^{n} \mathbb{E}\xi_{i_0 i_1} \xi_{i_1 i_2} ... \xi_{i_{k-1} i_0}$$

Due to the fact that entries are centered and independent some of the summands will be zeros and thus to proceed we need to calculate the number of summands that are non-zero which is a combinatorics problem. To handle it Vu used a counting scheme adapted from Furedi and Komlos. We had trouble understanding Vu's method, so split the task: David looked at the original method used by Furedi and Komlos, while I continued working on Vu's paper. Eventually we understood everything in the paper. For our summary of this work see: https://elidechse.github.io/VanVu.pdf.

Vu (2007, remark 1.5) commented: *"For the case when the entries of A are i.i.d. symmetric random variables, there are sharper bounds. The best current bound we know of is due to Soshnikov (1999), which shows that the error term [...] can be reduced to* $n^{-1/6+o(1)}$*."*

Clearly, it's a vast improvement of the bound but at the cost of an assumption of symmetry on the ξ_{ij}'s distributions.

As explained in section 1 it was the reference to Soshnikov (1999) that led me to the sequence of papers culminating in Péché et al. (2008):

- Sinai and Soshnikov (1998a) and Sinai and Soshnikov (1998b) proved a central limit theorem for traces of Wigner ensembles of symmetric random matrices with independent symmetrically distributed entries and developed the technique to count the number of summands.
- Soshnikov (1999) showed the convergence of the k largest eigenvalues of Wigner symmetric random matrices with independent symmetrically distributed sub-gaussian entries to the Tracy-Widom distribution.
- Soshnikov (2004) studied the behavior of the largest eigenvalue of a matrix with entries distribution having heavy tails.
- Soshnikov and Fyodorov (2005) studied a rectangular matrix with Cauchy entries and showed that its largest eigenvalues distribution does not follow Tracy-Widom law.
- Péché and Soshnikov (2007) showed that the trace expectation of a power of a matrix is close to the trace expectation of the standard Wigner matrix, leading to:

  $$\lambda(A) \leq 2\sigma\sqrt{n} + o(n^{-1/22+\varepsilon}),$$

  with high probability, where ε is an arbitrary small positive number.
- Péché et al. (2008) proved a lower bound for a random symmetric matrix with independent identically distributed entries.

We decided we had to start from the very beginning and tried to read the original paper (Sinai and Soshnikov (1998a)) and the accompanied one (Sinai and Soshnikov (1998b)). We had a trouble with one step of the argument (Sinai and Soshnikov, 1998a, (2.3)), and eventually I came up with a counterexmaple to this assertion. For details see: https://elidechse.github.io/Soshnikov.pdf

3 RDS

In many social and health studies it is difficult to reach the population of interest because of the nature of the study: if it is related to such aspects as drugs, sex, sexual orientation, HIV that are either very personal or stigmatized in the society then usual methods of sampling do not work any more. What researchers use instead is a method called Respondent Driven Sampling (RDS): they need to reach only several people of interest and then use them to recruit new participants for the study among their acquaintances.

In practice (Salganik and Heckathorn (2004)) everything starts with several initial participants called *seeds*. Then they recruit new participants among their acquintances through a system of unique coupons: when a new participant comes to the research facility with a coupon then the person who refered him is being paid, then after surveying the new participant gets his coupons and the process goes on until reaches the sample size. As human beings tend to cheat we cannot allow sampling with replacement: as all recruitments are paid for then it's too tempting to collaborate with someone to get recruited as much as possible, which would make the sample far from the desired random walk.

The method was introduced by Heckathorn (1997) in a paper that has been cited almost 3000 times. Subsequently, Volz and Heckathorn (2008) proposed an estimator that now is widely used in studies about HIV (Szwarcwald et al. (2011), Johnston et al. (2010)), racial studies (Wejnert (2010)), studies involving injection drug users (Wejnert et al. (2012)), etc

The most important assumption used in papers is that the sample collected as above is a random walk on the hidden population network, that is we assume that the sampling is with replacement and the recruitment is done uniformly at random among one's friends. The first assumption is always violated by the above note, but Volz and Heckathorn (2008) claim that with small sample sizes sampling with replacement is the same as sampling without replacement and thus researchers believe that the second assumption is met.

Unfortunately there is a huge gap between proved theoretical results and results that are hoped for and believed in in practice. Practitioners often act as if the seeds are drawn according to the stationary distribution and that even with a sample of a small size compared to the population size the estimator is close to the truth (Salganik and Heckathorn (2004)).

They act as if the whole recruitment scheme is an irreducible Markov chain, which starts from a stationary distribution. This Markov chain is shown to have a stationary distribution vector (which is also a vector of inclusion probabilities of nodes) proportional to the reciprocals on node degrees. These assumptions are ususally violated.

Goel and Salganik (2010) showed by simulation that in practice the behavior of Volz-Heckathorn estimator is worse than hoped for.

Guntuboyina et al. (2012) gave an example when assuming differential sampling node degree distributions were very different, thus making it impossible to recover the inclusion probabilities based solely on the degree distribution.

Rohe (2015) was actually considering a recruitment scheme when each participant recruits m of his friends and proved the convergence rates of variance of the estimator under different conditions, a theoretical result that I think is not particularly relevant to practice.

With simulations I compared the behavior of the VH estimator in ideal circumstances on different graphs: it's always close to the truth and is never worse than other estimators I tried. So the question is: can we get something with similar good behavior but with working assumptions?

References

Goel, S. and M. J. Salganik (2010). Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences 107*(15), 6743–6747.

Guntuboyina, A., R. Barbour, and R. Heimer (2012). On the impossibility of constructing good population mean estimators in a realistic respondent driven sampling model. *arXiv preprint arXiv:1209.2072*.

Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems 44*(2), 174–199.

Johnston, L., H. O'Bra, M. Chopra, C. Mathews, L. Townsend, K. Sabin, M. Tomlinson, and C. Kendall (2010). The associations of voluntary counseling and testing acceptance and the perceived likelihood of being HIV-infected among men with multiple sex partners in a south african township. *AIDS and Behavior 14*(4), 922–931.

Kyng, R., A. Rao, S. Sachdeva, and D. A. Spielman (2015). Algorithms for Lipschitz learning on graphs. In *COLT*, pp. 1190–1223.

Péché, S. and A. Soshnikov (2007). Wigner random matrices with non-symmetrically distributed entries. *Journal of Statistical Physics 129*(5), 857–884.

Péché, S., A. Soshnikov, et al. (2008). On the lower bound of the spectral norm of symmetric random matrices with independent entries. *Electronic Communications in Probability 13*, 280–290.

Rohe, K. (2015). Network driven sampling; a critical threshold for design effects. *arXiv preprint arXiv:1505.05461*.

Salganik, M. J. and D. D. Heckathorn (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology 34*(1), 193–240.

Sinai, Y. and A. Soshnikov (1998a). Central limit theorem for traces of large random symmetric matrices with independent matrix elements. *Boletim da Sociedade Brasileira de Matemática-Bulletin/Brazilian Mathematical Society 29*(1), 1–24.

Sinai, Y. G. and A. B. Soshnikov (1998b). A refinement of Wigner's semicircle law in a neighborhood of the spectrum edge for random symmetric matrices. *Functional Analysis and Its Applications 32*(2), 114–131.

Soshnikov, A. (1999). Universality at the edge of the spectrum in Wigner random matrices. *Communications in Mathematical Physics 207*(3), 697–733.

Soshnikov, A. (2004). Poisson statistics for the largest eigenvalues of Wigner random matrices with heavy tails. *Electronic Communications in Probability 9*, 82–91.

Soshnikov, A. and Y. V. Fyodorov (2005). On the largest singular values of random matrices with independent Cauchy entries. *Journal of Mathematical Physics 46*(3), 033302.

Szwarcwald, C. L., P. R. B. de Souza Júnior, G. N. Damacena, A. B. Junior, and C. Kendall (2011). Analysis of data collected by rds among sex workers in 10 Brazilian cities, 2009: estimation of the prevalence of HIV, variance, and design effect. *Journal of Acquired Immune Deficiency Syndromes 57*, S129–S135.

Tao, T. (2012). *Topics in random matrix theory*, Volume 132. American Mathematical Society Providence, RI.

Volz, E. and D. D. Heckathorn (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics 24*(1), 79.

Vu, V. H. (2007). Spectral norm of random matrices. *Combinatorica 27*(6), 721–736.

Wejnert, C. (2010). Social network analysis with respondent-driven sampling data: A study of racial integration on campus. *Social Networks 32*(2), 112–124.

Wejnert, C., H. Pham, N. Krishna, B. Le, and E. DiNenno (2012). Estimating design effect and calculating sample size for respondent-driven sampling studies of injection drug users in the United States. *AIDS and Behavior 16*(4), 797–806.

Zhou, D., C. J. Burges, and T. Tao (2007). Transductive link spam detection. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pp. 21–28. ACM.