Prospectus

Elena Khusainova

26 April 2017

1 Introduction

During my second year I took two courses that influeced me most: in fall semester I took Dan Spielman's course on spectral graph theory and in 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. *Say anything about it?*

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 working on spectral theory more intensely. *attracted?*

As described in the section 2 we made a lot of progress, spent quite some time reading related literature, and eventually discovered that state of the art Peche and Soshnikov's bound: Péché and Soshnikov (2007). Unfortunately this paper is a part of a sequence of several papers written by Soshnikov and his coathors. They develop an elaborate technique to address various problems of the field, and the one we were interested in relies heavily on the first paper by Sinai and Soshnikov (1998a) in the sequence in which we found what appears to be a serious error. Reluctantly, decided to put it on hold as otherwise it would require sugnificant effort and time to correct the mistake in a way that doesn't disrupt the whole logic of the technique.

As I still wanted to work on networks, Yihong 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 Vu

One of the important questions in spectral theory is the behavior of the spectrum of a matrix and its largest eigenvalue in particular. That is consider a random $n \times n$ symmetric matrix A with independent entries ξ_{ij}. Denote $|\lambda_1(A)| \geq |\lambda_2(A)| \geq ... \geq |\lambda_n(A)|$ to be its eigenvalues. Then we are interested in the bound for $\lambda_1(A)$ as $n \to \infty$.

Vu (2007) studied a random matrix A with centered bounded entries with bounded variance:

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

To upper bound the spectral norm of A Vu used Wigner method that involves calculating the trace of powers of the matrix:

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

where k is even. If we take the expectation of both sides then:

$$\mathbb{E}\lambda_1(A)^k \leq \sum_{i_0=1}^{n} \sum_{i_1=1}^{n} ... \sum_{i_{k-1}=1}^{n} \mathbb{E}\xi_{i_0 i_1} \xi_{i_1 i_2} ... \xi_{i_{k-1} i_0} \quad (2)$$

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 the method, so split the task: David looked at the original method used by Furedi and Komlos, while I was going through Vu's explantion of his scheme, and eventually we figured it out and slightly improved. (For our cleaned version on the Van Vu's coding scheme see: https://elidechse.github.io/VanVu.pdf)

Why is it necessary to mention that we split the task?

Vu cited (Vu, 2007, remark 1.5) one of the Soshnikov papers (Soshnikov (1999)) noting that his was the best result of that time for the case of symmetrically distributed entries. I did a literature search and found a more recent paper Péché and Soshnikov (2007) that improved Vu's bound. In fact Soshnikov has written many papers on the topic:

- In Sinai and Soshnikov (1998a) and Sinai and Soshnikov (1998b) they 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.
- In Soshnikov (1999) they 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.
- In Péché and Soshnikov (2007) the key result that helped achieve the bound was proved: they showed that the trace expectation of a power of a matrix is close to the trace expectation of the standard Wigner matrix.

Additionally:

Check!

- In Soshnikov et al. (2004) they studied the behavior of the largest eigenvalue of a matrix with entries distribution having heavy tails.
- In Soshnikov and Fyodorov (2005) they studied a rectangular matrix with Cauchy entries and showed that its largest eigenvalues distribution does not follow Tracy-Widom law.
- In Peche et al. (2008), they proved a lower bound for a random symmetric matrix with independent identically distributed entries.
- etc.

More?

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 (for details see: https://elidechse.github.io/Soshnikov.pdf) to this assertion.

Should I explain here on anywhere more details about Soshnikov's big picture?

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 to 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. With 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) and now the paper has been cited almost 3000 times, showing huge interest in such methods. Then 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 replacment 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 replacment 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. In practice people believe that 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)).

On the other hand most theoretical results imply a recruitment scheme when each participant recruits only one of his friends and thus the whole recruitment sample is considered to be an irreducible Markov chain, that 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.

Goel and Salganik (2010) showed by simulation that in practice the behavior 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.

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.

Peche, S., A. Soshnikov, et al. (2008). On the lower bound of the spectral norm of symmetric random matrices with independent entries. *Electron. Commun. Probab 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. et al. (2004). Poisson statistics for the largest eigenvalues of wigner random matrices with heavy tails. *Electron. Comm. Probab 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. *JAIDS Journal of Acquired Immune Deficiency Syndromes 57*, S129–S135.

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.