Prospectus

Elena Khusainova

26 April 2017

1 Introduction

I've been working on several problem involving networks, graphs, random matrices, inspired by the Van Vu's and Dan Spielman's courses I was taking during my second year.

With Harry I started working on the problem of spam detection based on Kyng et al. (2015) and Zhou et al. (2007). At the same tine I was exploring general network theory (Davis-Kahan, random matrix theory, spectral theory).

In particular reading Vu (2007) on behavior of spectral norm of a random matrix.

I made some progress with spam but found myself more interested in the underlying theory. Started working on spectral theory more intensely.

As described in the section 2 we made a lot of progress, read a lot of literature, eventually discovered that state of art Soshnikov. Unfortunately Ssoshnikov's papers rely heavily on the first paper by Sinai and Soshnikov (1998a) in the sequence. We found what appear to be a serious error. Reluctantly, decided to put it on hold.

Still wanted to work on networks. Yihong suggested a problem in RDS, a technique that appears to be much used in the social science literature despite the lack of theory. Since February I have been exploring with simulations which appear promising. I want to concentrate on this.

2 Vu

Vu (2005) studied random $n \times n$ symmetric matrix A with independent bounded entries ξ_{ij} with bounded variance:

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

Vu uses Wigner method that involves calculating the trace of powers of the matrix:

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

where k is even.

Need to calculate number of summands 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. We figured it out. (`run:file:///Users/elenakhusainova/Documents/Study/Webpage/VanVu.pdf`My work on Van Vu)

Vu cited (Vu, 2007, remark 1.5) one of the Sinai and Soshnikov papers (Soshnikov (1999)) noting that theirs 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 that improved the Vu's bound. In fact Soshnikov has written lots of papers on the topic:

- In Sinai and Soshnikov (1998a) and Sinai and Soshnikov (1998b) they prve CLT for Wigner (symmetric with symetrically distributed entries?) matrices and developed the technique to count the number of summands

- In Soshnikov (1999) they proved the behaviour of the the k largest eigenvalues of a Wigner matrix
- 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:

- In Soshnikov et al. (2004) they studies the behaviour of the largest eigenvalue of a matrix with entries distribution having heavt 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 **?**, they proved a lower bound for a random symmetric matrix with iid entries.
- etc.

We decided we needed to go to 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)), then I eventually came up with a counterexmaple (see online) to this assertion.

3 RDS

In some 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.

The method was introduced by Heckathorn in Heckathorn (1997) and now the paper has been cited almost 3000 times. The extimatro proposed by Volz and Heckathorn in Heckathorn (2007) is used in studies about HIV (Szwarcwald et al. (2011), Johnston et al. (2010)), racial studies (Wejnert (2010)), injection drug users (Wejnert et al. (2012)), etc

, similar to random walk on a graph and all the theoretical results (none?) were done as if it's a random walk or Markov Chain, but in practice we cannot allow the walk to come back to the same vertex twice as human beings would be tempted to trick the system to earn money. Thus it's not exactly a MC. It has been shown (Goel and Salganik (2010), Rohe (2015), Guntuboyina et al. (2012)) that violation of assumptions might lead to bad results. So random walk is oversimplification. But the simulations showed that under the assumptions the proposed (RDS II) estimator is quite good and promising, so the question is: can 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.

Heckathorn, D. D. (2007). Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology 37*(1), 151–207.

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.

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

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.

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.