Graphons: A Nonparametric Method to Model, Estimate, and Design Algorithms for Massive Networks[*]

CHRISTIAN BORGS, Microsoft Research
JENNIFER CHAYES, Microsoft Research

Many social and economic systems are naturally represented as networks, from off-line and on-line social networks, to bipartite networks, like Netflix and Amazon, between consumers and products. Graphons, developed as limits of graphs, form a natural, nonparametric method to describe and estimate large networks like Facebook and LinkedIn. Here we describe the development of the theory of graphons, for both dense and sparse networks, over the last decade. We also review theorems showing that we can consistently estimate graphons from massive networks in a wide variety of models. Finally, we show how to use graphons to estimate missing links in a sparse network, which has applications from estimating social and information networks in development economics, to rigorously and efficiently doing collaborative filtering with applications to movie recommendations in Netflix and product suggestions in Amazon.

Additional Key Words and Phrases: Networks; graphons; graph convergence; nonparametric estimation; collaborative filtering; network completion

Networks are the natural framework to describe social and economic systems with pairwise interactions. This includes not only social networks among individuals, like Facebook and LinkedIn, but also bipartite networks between different types of entities, like the Netflix movie-user network and the Amazon consumer-product network. Graphons, developed as limits of sequences of graphs, form a natural, non-parametric method to model and estimate large networks. Here we informally describe some of the development of the theory of graphons over the last dozen years, and provide an application of graphons to the completion of noisily, sparsely sampled massive networks.

1 GRAPH LIMITS

The theory of limits of sequences of dense graphs (i.e., graphs in which the number of edges scale as the square of the number of vertices) was developed over roughly a decade starting around 2006 [19–21, 37]. Of course, it is always possible to a define a limit of a graph sequence – if we make the notion too weak, then every sequence trivially converges to the same point; if we make it too strong, each sequence could converge to a different point. The key was to find a notion that was "just right." This was done in a series of papers with Lovász, Sos and Vestergombi [20, 21].

In [19] we define a metric to allow us to compare graphs of different sizes. For this, we rescale two graphs to be of the same size by creating copies of each vertex so that the two graphs then have
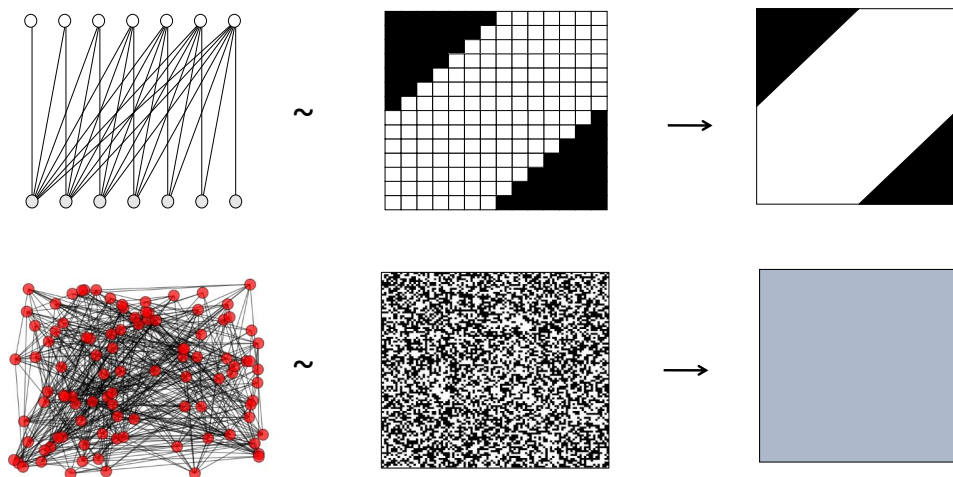

Fig. 1. The half-graph and a random graph, their empirical graphons, and their limits.

the same number of vertices. Specifically, if the two graphs have n_1 and n_2 vertices, and k_1 and k_2 are the smallest numbers such that $k_1 n_1 = k_2 n_2$, we "blow up" each vertex in the first graph by k_1, and each one in the second by k_2. We then replace each edge in the original graphs by complete bipartite graphs between the blown-up vertices. We look at the Frieze-Kannan cut distance [29] of the difference between the two blown-up graphs, where the cut distance is basically the value of the cut between a set and its complement, taking the sup over all sets. Finally, since vertex labels should not matter, we take the minimum over all relabelings of the vertices. The resulting metric is called the cut metric.

Our first notion of convergence [19, 20] is what we call metric convergence, which means simply that the sequence of graphs is Cauchy in the cut metric. The fact that a limit exists was established by Lovász and Szegedy [37] using a weak version of the Szemerédi Regularity Lemma [29] and a martingale argument. The limit is a two-variable symmetric function, called a graphon, often denoted by $W(x, y)$, which is basically a continuous version of the adjacency matrix of a graph on $[0, 1]^2$. In [12], we show that the limit is unique up to measure-preserving transformations of the underlying feature space.

Note that a graph on n vertices can be represented as a piecewise constant graphon by dividing the unit square $[0, 1]^2$ into n^2 small squares of side length $1/n$, and representing the presence of an edge between vertices i and j by a 1 on the squares ij and ji, and its absence by a 0. This representation is called the empirical graphon of the graph. If we visualize a 0 in the empirical graphon as a small white square, and a 1 as a small black square, graph convergence can be seen as convergence of the these black and white empirical graphons to a gray-scale limiting graphon. See Figure 1 for an illustration for the half-graph and an Erdös-Renyí random graph.

There is a natural local notion of convergence in dense graphs. We can ask that all induced finite subgraph densities (e.g., the edge density, triangle density, Peterson graph density, ...) converge simultaneously. We prove that a sequence of dense graphs converges in metric if and only if it is subgraph convergent [20].

We also introduce several "global" notions of convergence [10, 21]. We have one notion which corresponds to convergence of all finite two-body statistical physics models in the microcanonical ensemble (the ensemble in which each species of vertex represents a fixed percentage of all vertices). We have another notion which corresponds to all multi-way cuts and bisections converging. We show that a sequence of dense graphs converges according to these and a few other global notions if and only if it converges in metric [21]. More recently, a notion of large deviations convergence was defined [10], which is also equivalent with the other global and local notions for dense graphs.

Sparse graphs of bounded degree are much more fragile. Here we have no notion of metric convergence. The natural notion of local convergence is convergence of a graph rooted at a random point, known as Benjamini-Schramm convergence [5]. Benjamini-Schramm convergence is equivalent to subgraph convergence [11], but global notions of convergence are stronger [10].

The most interesting case is that of sparse graphs of unbounded average degree, like the power-law networks in Facebook and LinkedIn. These cases were much more complicated due to the fact that much of the necessary graph theory (e.g., the relevant case of the weak Regularity Lemma) had not yet been developed. Over the past five years, we have developed two ways to generalize the theory.

The first way is by rescaling the empirical graphon function, dividing it by the edge density ρ of the graph sequence (which is tending to zero in the case of sparse graphs). This was done by Bollobás and Riordan in the case when the graphs had no "dense spots," corresponding to the graphon being a bounded function [8]. In collaboration with Cohn and Zhao [15, 16], we develop a theory of "uniformly upper regular" sparse graphs of unbounded average degree; we show that these converge to graphons in L^p spaces, corresponding to growing sequences with power-law tails. In the process of doing this, we formulate and prove a weak form of the Regularity Lemma for sparse graphs corresponding to graphons of unbounded degree.

The second way to get graph limits for graphs of unbounded average degree, again graphs with very long tails, is by "stretching" the empirical graphon (and its domain) by $1/\sqrt{\rho}$ so that the integral of the graphon is again unity. This leads to a limit defined on the positive quadrant rather than the unit square, and hence these graphons provide a good description of networks where the underlying features cannot be restricted to bounded domains. See Figure 2 for a comparison of this with the rescaling approach. In work with Cohn and Holden [17], we develop a theory of sparse graphs of unbounded degree with "uniformly regular tails," show how to construct limits for sequences obeying this regularity condition, and prove many properties of the these graphons. Finally, we show that any graph sequence which satisfies both the conditions of uniform upper regularity and uniformly regular tails must be dense; so dense graphs are the only overlap of our two theories of sparse graphs of unbounded degree.

2 MODELING LARGE NETWORKS

The classic way to model large networks uses the so-called stochastic block model [32], which is a generalization of the Erdös-Renyí random graph to k species of vertices, with fixed densities of each species of vertex, and fixed probabilities that any two species connect to each other, including a fixed probability of intra-species connection, which is in general different within each species.

As graphs get larger and larger, reasonable descriptions have the number of species scale with the number of vertices: $k = k(n)$, which tends to lead to overfitting. Instead, starting over a decade ago [31], researchers began to model large networks non-parametrically by choosing n points randomly from some feature space, $x_1, x_2, \cdots, x_n \in \Omega$, and a two-variable, bounded symmetric function $W(x_i, x_j) : \Omega \times \Omega \to [0, 1]$. A random graph consistent with this function is then generated by saying that i is connected to j with probability $W(x_i, x_j)$.
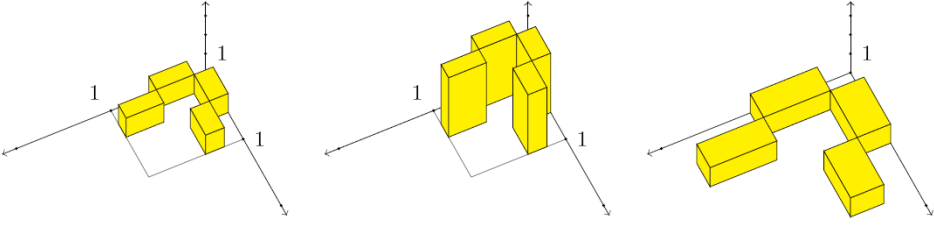
Fig. 2. Three graphons associated with the same simple graph G on five vertices: the emprical graphon used in theory of dense graph convergence, the rescaled a version used in the uniformly upper regular theory of sparse graph convergence, and the stretched version used in the theory for sparse graphs with uniformly regular tails.

The graphs generated in this way can easily be seen to be dense, whereas real life networks tend to be sparse. But our two methods for constructing graph limits for sequences of random sparse graphs of unbounded degree naturally yield random models of massive sparse networks.

Our first method [15] chooses features $x_1, x_2, \cdots, x_n$ from some compact feature space, and generates random graphs with long tails corresponding to unbounded graphons W, connecting vertices with features x_i and x_j with probability $\min\{1, \rho W(x_i, x_j)\}$, where ρ is the desired target density of our graph.

Our second method [17] is more complex: It chooses features $x_1, x_2, \cdots$ from an unbounded feature space according to a Poisson process with an intensity which scales linearly with time. At time t, the process connects vertices i and j born up to time t with probability $W(x_i, x_j)$. The resulting graph at time t retains only those vertices which have been connected to at least one other vertex by that time, leading to a graph with a finite number of vertices at any finite time. We show that this kind of model is quite natural by proving a generalization of the celebrated Aldous-Hoover Theorem [3, 33] to sparse graphs. In our setting, it says that under suitable regularity conditions, any family of graphs with vertices labeled by their birth time that is invariant under measure-preserving transformations of time must be of the above form. Our model generalizes the work of Caron and Fox [22], and is very similar to a model introduced simultaneously by Veitch and Roy [41].

3 LEARNING SPARSE GRAPHS

How does one learn a network from a single sample, e.g., allowing us to predict how the network might look when it's twice its current size? The traditional approach is to assume that the network is a stochastic block model with k types of vertices, cluster the vertices according to type, and then use this clustering to estimate the parameters of the model (the densities of each of the k types, plus the $k(k+1)/2$ connection probabilities within and between types) and then use that estimated stochastic block model to make predictions about the network [4, 9, 24, 26–28, 32, 34, 38, 40, 42, 43]. See [1, 2] for recent work with rigorous consistency guarantees for sparse graphs.

As we discussed above, for very large networks, it is often better to take a non-parametric approach [6, 7, 25, 35, 39]. Here the idea is to get an estimate, $\widehat{W}$ for the graphon W representing the network, and then generate realizations from that estimation. For bounded graphons, it had been shown that this leads to consistent estimation for both dense and sparse graphs, see, e.g., [23, 30, 36, 44], to name a few of the more recent papers on the subject. In collaboration with

Cohn and Ganguly [14], we showed that this leads to consistent estimation for sparse massive networks if, e.g., $W \in L^2$, provided that the average degree diverges. Hence one can use graphons for statistically consistent estimation for power-law graphs. In collaboration with Smith [13], we showed that, for sparse graphs corresponding to bounded W, it is possible to consistently estimate while maintaining edge differential privacy.

4 NETWORK COMPLETION AND COLLABORATIVE FILTERING ON SPARSE GRAPHS

An area of some interest in economics is how to "complete" sparsely, noisily sampled networks. For example, in development economics, one sometimes has measurements representing sparsely and noisily sampled connections among individuals. How sparsely can we sample and still get an accurate estimate of the missing links? Similarly, in bipartite networks like the Netflix user-movie network or the Amazon consumer-product network, can we show that we can complete the network even if it is very sparsely sampled with some general noise? Here we follow a collaborative filtering approach, which however must be modified to account for the sparsity of the sample. What if the sampling is so sparse that two people for whom we are trying to estimate the probability of a connection have no connections in common? Or what if there are no sets of movies in common between a given individual and another on the Netflix network? In collaboration with Lee and Shah [18], we used the "expanded neighborhood" approach of Abbe and Sandon [1, 2], who study estimation for stochastic block models. The idea is to look not just at individuals who share a common connection, but instead expand the neighborhood until a common connection is found. Similarly, we expand the bipartite user-movie-user-movie-$\cdots$ neighborhood until there an overlap. Here we treat graphons as operators in the expansion and need to do estimates to control the variance. Assume there are n individuals on the network, each described by features in a d-dimensional space. We show that, if the underlying graphon describing this network is Lipshitz, then the mean square error of our estimated graphon tends to zero provided that we sample the network with density $p = \omega(d^2 n^{-1})$. For high-dimensional latent spaces, which tend to be the appropriate descriptions for most real-world networks, this is a significant improvement over the best previous result, that of Chatterjee [23], who showed that the mean square error of the estimated graphon tends to zero provided that the network is sampled with density $p = \omega(n^{-2/(d+2)})$.

REFERENCES

[1] E. Abbe and C. Sandon. 2015. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. In *Proceedings of the 56th Annual Symposium on Foundations of Computer Science (FOCS'15)*. 670–688.

[2] E. Abbe and C. Sandon. 2015. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in Neural Information Processing Systems 28 (NIPS'15)*. 676–684.

[3] D. Aldous. 1981. Representations for partially exchangeable arrays of random variables. *J. Multivar. Anal.* 11 (1981), 581–598.

[4] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. 2014. A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research* 15 (2014), 2239–2312.

[5] I Benjamini and O Schramm. 2001. Recurrence of distributional limits of finite planar graphs. *Electronic Journal of Probability* 23 (2001), 1–13.

[6] P. J. Bickel and A. Chen. 2009. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America* 106 (2009), 21068–21073.

[7] P. J. Bickel, A. Chen, and E. Levina. 2011. The method of moments and degree distributions for network models. *Annals of Statistics* 39 (2011), 2280–2301.

[8] B. Bollobás and O. Riordan. 2009. Metrics for sparse graphs. In *Surveys in combinatorics 2009 (eds. S. Huczynska, J. D. Mitchell, and C. M. Roney-Dougal), London Math. Soc. Lecture Note Ser. **365**, Cambridge University Press*. London Math. Soc. Lecture Note Ser. **365**, Cambridge University Press, 211–287.

[9] R. B. Boppana. 1987. Eigenvalues and graph bisection: an average-case analysis. In *28th Annual Symposium on Foundations of Computer Science.* 280–285.

[10] C. Borgs, J. Chayes, and D Gamarnik. 2013. Convergent sequences of sparse graphs: A large deviations approach. *preprint, arXiv:1302.4615. To appear in Random Structures Algorithms* (2013).

[11] C. Borgs, J. Chayes, J Kahn, and L. Lovász. 2013. Left and right convergence of graphs with bounded degree. *Random Structures Algorithms* 42 (2013), 1–28.

[12] C. Borgs, J. Chayes, and L. Lovász. 2010. Moments of two-variable functions and the uniqueness of graph limits. *Geometric And Functional Analysis* 19, 6 (2010), 1597–1619.

[13] C. Borgs, J.T. Chayes, and A. Smith. 2015. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems 28.* 1369–1377.

[14] C Borgs, J. T. Chayes, H Cohn, and S Ganguly. 2015. Consistent nonparametric estimation for heavy-tailed sparse graphs. *preprint, arXiv:1508.6675* (2015).

[15] C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao. 2014. An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. *preprint, arXiv:1401.2906* (2014).

[16] C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao. 2014. An L^p theory of sparse graph convergence II: LD convergence, quotients, and right convergence. *preprint, arXiv:1408.0744* (2014).

[17] C Borgs, J. T. Chayes, and N Cohn, H andd Holden. 2016. Sparse exchangeable graphs and their limits via graphon processes. *preprint, arXiv:1601.07134. To appear in Journal of Machine Learning Research* (2016).

[18] C Borgs, J. T. Chayes, C. E Lee, and D. Shah. 2017. Sparse graphon Estimation via Neighbor Method. *preprint* (2017).

[19] C. Borgs, J. T. Chayes, L. Lovász, V. Sós, and K. Vesztergombi. 2006. Counting graph homomorphisms. In *Topics in discrete mathematics (eds. M. Klazar, J. Kratochvíl, M. Loebl, J. Matoušek, R. Thomas, and P. Valtr), Springer.* Springer, 315–371.

[20] C. Borgs, J. T. Chayes, L. Lovász, V. Sós, and K. Vesztergombi. 2008. Convergent graph sequences I: subgraph frequencies, metric properties, and testing. *Advances in Math.* 219 (2008), 1801–1851.

[21] C. Borgs, J. T. Chayes, L. Lovász, V. Sós, and K. Vesztergombi. 2012. Convergent graph sequences II: multiway cuts and statistical physics. *Ann. of Math.* 176 (2012), 151–219.

[22] F. Caron and E. B Fox. 2014. Sparse graphs using exchangeable random measures. *preprint, arXiv:1401.1137* (2014).

[23] S. Chatterjee. 2015. Matrix estimation by universal singular value thresholding. *Annals of Statistics* 43, 1 (2015), 177–214.

[24] Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. 2012. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Proceedings of the 25th Annual Conference on Learning Theory (JMLR Workshop and Conference Proceedings Volume 23).* 35.1–35.23. http://www.jmlr.org/proceedings/papers/v23/chaudhuri12/chaudhuri12.pdf

[25] D. S. Choi, P. J. Wolfe, and E. M. Airoldi. 2012. Stochastic blockmodels with a growing number of classes. *Biometrika* 99 (2012), 273–284.

[26] A. Coja-Oghlan. 2010. Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.* 19 (2010), 227–284.

[27] A. Dasgupta, J. Hopcroft, R. Kannan, and P. Mitra. 2006. Spectral clustering by recursive partitioning. In *Algorithms – ESA 2006 (Lecture Notes in Computer Science),* Y. Azar and T. Erlebach (Eds.), Vol. 4168. Springer, 256–267. http://dx.doi.org/10.1007/11841036_25

[28] S.E. Fienberg, M.M. Meyer, and S.S. Wasserman. 1985. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association* 80 (1985), 51–67.

[29] A. Frieze and R. Kannan. 1999. Quick approximation to matrices and applications. *Combinatorica* 19 (1999), 175–220.

[30] C. Gao, Y. Lu, and H. H. Zhou. 2015. Rate-optimal graphon estimation. *Annals of Statistics* 43 (2015), 2624–2652.

[31] P. D. Hoff, A. E. Raftery, and M. S. Handcock. 2002. Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* 97, 460 (2002), 1090–1098.

[32] P.W. Holland, K.B. Laskey, and S. Leinhardt. 1983. Stochastic blockmodels: first steps. *Social Networks* 5 (1983), 109–137.

[33] D. Hoover. 1979. Relations on probability spaces and arrays of random variables. *preprint, Institute for Advanced Study, Princeton, NJ* (1979).

[34] Mark Jerrum and Gregory B. Sorkin. 1998. The Metropolis algorithm for graph bisection. *Discrete Appl. Math.* 82 (1998), 155–175.

[35] O. Kallenberg. 1999. Multivariate sampling and the estimation problem for exchangeable arrays. *Journal of Theoretical Probability* 12 (1999), 859–883.

[36] Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. 2017. Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* 45 (2017), 316–354.

[37] L. Lovász and B. Szegedy. 2006. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B* 96 (2006), 933–957.

[38] F. McSherry. 2001. Spectral partitioning of random graphs. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*. 529–537.

[39] K. Rohe, S. Chatterjee, and B. Yu. 2011. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* 39 (2011), 1878–1915.

[40] T. A. B. Snijders and K. Nowicki. 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification* 14 (1997), 75–100.

[41] V. Veitch and D. M Roy. 2015. The class of random graphs arising from exchangeable random measures. *preprint, arXiv:1512.03099* (2015).

[42] Yuchung J. Wang and George Y. Wong. 1987. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association* 82 (1987), 8–19.

[43] H.C. White, S.A. Boorman, and R.L. Breiger. 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology* 80 (1976), 730fi??780.

[44] P. J. Wolfe and S. C. Olhede. 2013. Nonparametric graphon estimation. *preprint, arXiv:1309.5936* (2013).