Graphons, mergeons, and so on!

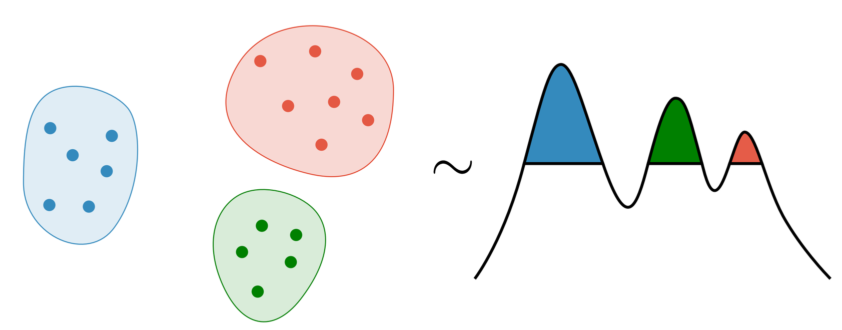
Justin Eldridge, Mikhail Belkin, Yusu Wang

The Ohio State University

We develop a statistical theory of graph clustering, where the goal is to organize the nodes of a graph into clusters. We model graphs as being generated from a very powerful random graph model of much recent interest called a *graphon*. One interpretation of the graphon is as an infinite graph. Sampling a graph from a graphon is analogous to inducing a random subgraph. We define the clusters of the graphon; naturally, the goal of clustering within the model is the recovery of these clusters. We develop a notion of convergence to the clusters of the graphon, and prove that a practical clustering algorithm is convergent; i.e., statistically consistent.

1) A statistical approach to clustering.

In this view, *clustering* is statistical estimation.

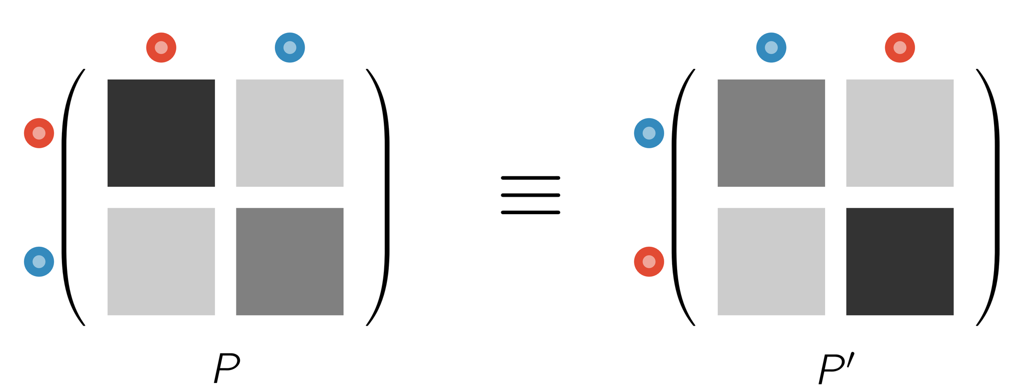


Steps to a statistical theory of clustering:

0. Model the data as a sample from a distribution.
1. Define the clusters of the distribution.
2. Develop a notion of statistical consistency.
3. Construct consistent algorithms.

2) Background: stochastic blockmodels.

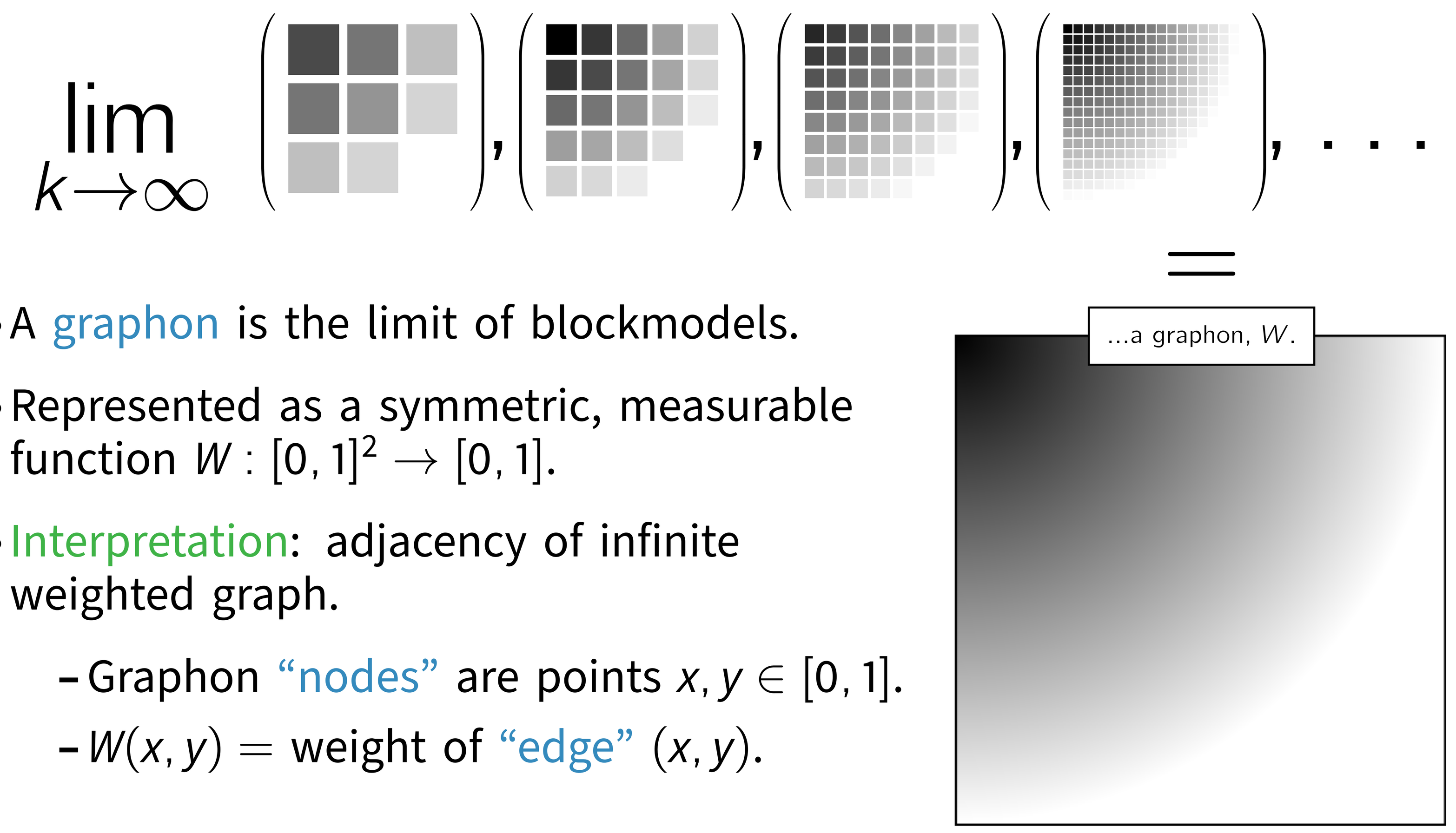
- Much existing graph clustering theory assumes that graphs are generated by the stochastic blockmodel.
- Each node belongs to one of k blocks or communities.
- Edge probabilities parameterized by $k \times k$ matrix P.
  - P_{ij} = probability of edge between communities i and j.
- Permutations of P do not change graph distribution.
- Example, 2-blockmodel:



- Consistency: Recover the correct community label of each node in unlabeled graph.
- Spectral methods like (McSherry, 2001) are consistent.

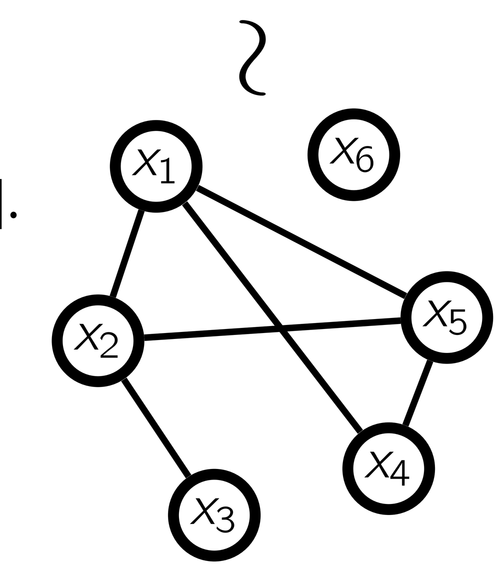
3) The graphon model.

- Problem: Many real-world networks are not well-modeled by the simple 2-blockmodel.
- Idea: Increase the number of blocks; (Lovász, 2012).



- A graphon is the limit of blockmodels.
- Represented as a symmetric, measurable function $W : [0,1]^2 \to [0,1]$.
- Interpretation: adjacency of infinite weighted graph.
  - Graphon "nodes" are points $x, y \in [0,1]$.
  - $W(x,y)$ = weight of "edge" (x,y).

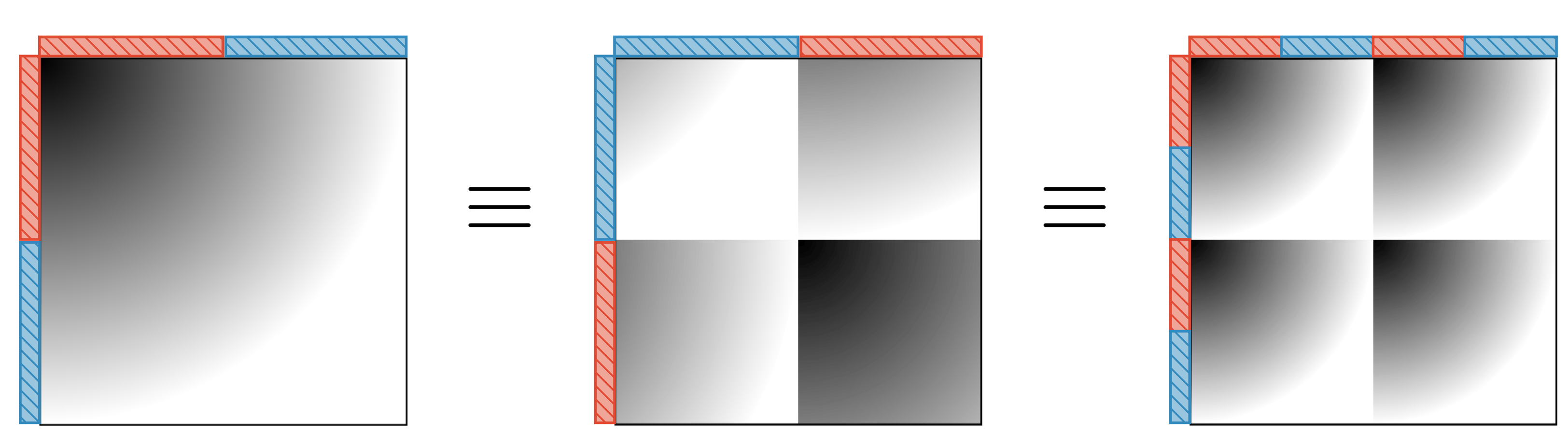
Sampling a graph from a graphon.



- To sample a graph of n nodes from W:
  1. Draw n graphon nodes $x_1, \ldots, x_n \sim \text{Unif}[0,1]$.
  2. Add edge (x_i, x_j) with probability $W(x_i, x_j)$.
- Much richer model than simple blockmodels.

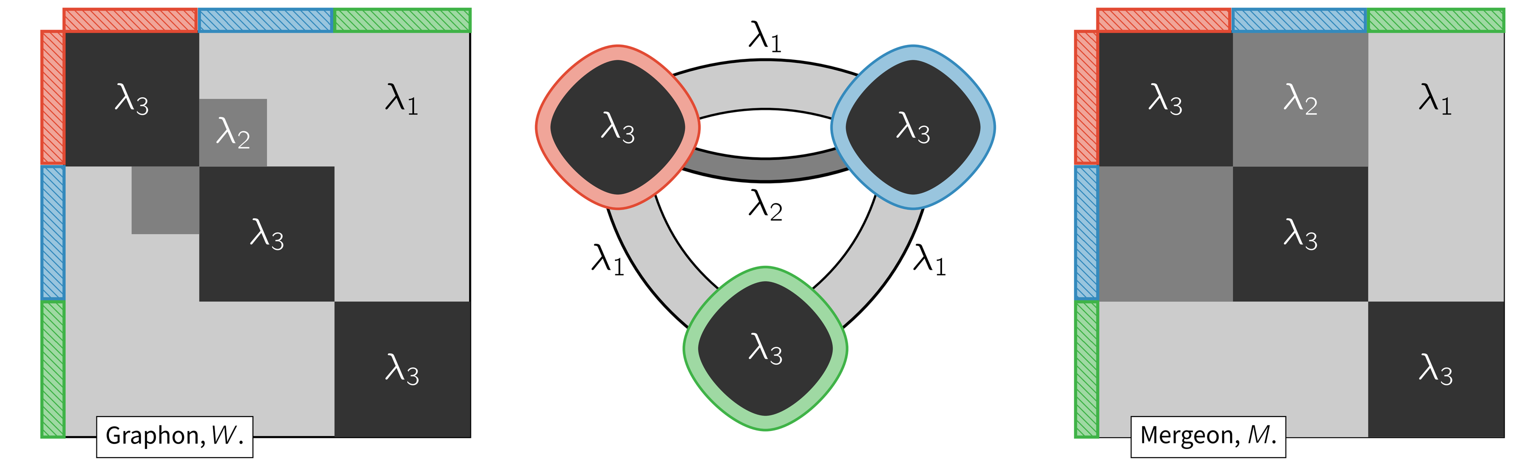
Two subtleties of infinite graphs.

1. A single graphon node/edge is insignificant.
   - That is, $W_1 \overset{\text{a.e.}}{=} W_2$ are equivalent.
   - Right way: Deal with equivalence classes of sets modulo null sets.
2. Graphon relabelings can be very complex.
   - Graphons W_1 and W_2 define the same graph distribution $\iff$ if they are the same up to relabeling $\varphi : [0,1] \to [0,1]$.
   - Only condition on φ: measure of $\varphi^{-1}(C)$ = measure of C.
   - Can map null set to set of full measure; very far from a bijection.

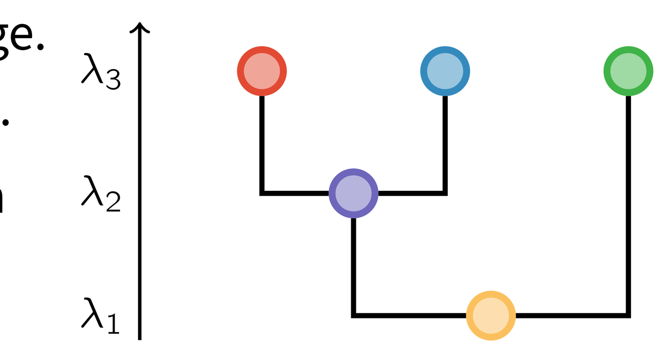


4) The clusters of a graphon.

- Interpret a graphon as adjacency of infinite weighted graph.

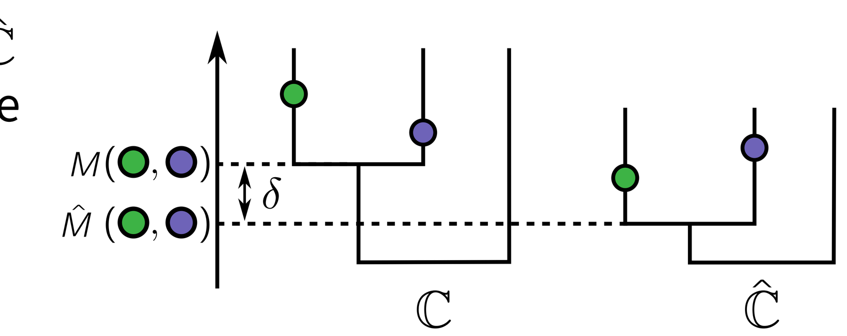


- Use generalized notion of connectivity, extending (Janson, 2008):
  - A set A is disconnected at level λ if it can be partitioned into S and $A \setminus S$ such that $W < \lambda$ a.e. on $S \times (A \setminus S)$. Otherwise, connected at level λ.
- Define clusters as connected components at each level.
  - At level λ_3: red, blue, and green regions are separate clusters.
  - At level λ_2: red and blue clusters merge.
  - At level λ_1: all clusters merge into one.
- Encode first height at which nodes x, y in same cluster with $M(x,y)$.
- We call $M : [0,1]^2 \to [0,1]$ a mergeon.
- Mergeon has hierarchical structure: call this the graphon cluster tree.



5) Consistency in merge distortion.

- We formalize the sense in which the output $\hat{\mathbb{C}}$ of a hierarchical clustering algorithm converges to graphon cluster tree $\mathbb{C}$.
- Measure distance between $\mathbb{C}$ and $\hat{\mathbb{C}}$ by comparing height at which same pair of nodes merge in each tree.
- Use the mergeons M and $\hat{M}$ for $\mathbb{C}$ and $\hat{\mathbb{C}}$ to define merge distortion: $d(\mathbb{C}, \hat{\mathbb{C}}) = \max_{\bullet,\bullet} |M(\bullet,\bullet) - \hat{M}(\bullet,\bullet)|$.



- Consistency: $d(\mathbb{C}, \hat{\mathbb{C}}) \to 0$ w.h.p. as number of nodes $n \to \infty$.

6) Consistent graph clustering algorithms.

- Intuitively, edge (i,j) has latent probability P_{ij}, induced by W.
- Theorem: If edge probability estimator $\hat{P}$ consistent in max-norm, i.e., $\max_{i \neq j} |P_{ij} - \hat{P}_{ij}| \to 0$ w.h.p. as $n \to \infty$, single-linkage on $\hat{P}$ is consistent.
- Problem: No such results for existing edge probability estimators. We modify and analyze neighborhood smoothing, (Zhang et al., 2015):
- Theorem: Modified nbhd. smoothing estimator is consistent in max norm.
- Corollary: Modified nbhd. smoothing + single linkage is a consistent clustering algorithm in the graphon model.