Property testing, parameter estimation, and graph limits

Jan Hladký
Institute of Mathematics
Academy of Sciences of the Czech Republic

JH's research is supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme.

1866 Gregor Johann Mendel

"factors" (=genes), and there are certain laws of inheritance

1866 Gregor Johann Mendel

"factors" (=genes), and there are certain laws of inheritance

1944	Avery-MacLeod-McCarty	*DNA is a double helix carrying the genetic information*
1952	Hershey-Chase	
1953	Watson-Crick	
~ 1959	Nirenberg, ...	

1866 Gregor Johann Mendel

"factors" (=genes), and there are certain laws of inheritance

Genes are stored on a linear structure! Why?

1944	Avery-MacLeod-McCarty
1952	Hershey-Chase
1953	Watson-Crick
$\sim$ 1959	Nirenberg, ...

DNA is a double helix carrying the genetic information

1866 Gregor Johann Mendel

"factors" (=genes), and there are certain laws of inheritance

Genes are stored on a linear structure! Why?

Compute the covariances of the gene switches.

Example: offsprings of a unisexual organism

○	*red*	↑	*little*
○	*red*	↓	*little*
□	*blue*	↑	*little*

□	*blue*	↓	*little*
○	*red*	↓	*BIG*
○	*red*	↑	*little*

1944	Avery-MacLeod-McCarty	*DNA is a double helix carrying the genetic information*
1952	Hershey-Chase	
1953	Watson-Crick	
$\sim$ 1959	Nirenberg, ...	

1866 Gregor Johann Mendel

"factors" (=genes), and there are certain laws of inheritance

Genes are stored on a linear structure! Why?

Compute the covariances of the gene switches.

Example: offsprings of a unisexual organism

○	*red*	$\uparrow$	*little*
○	*red*	$\downarrow$	*little*
□	*blue*	$\uparrow$	*little*

□	*blue*	$\downarrow$	*little*
○	*red*	$\downarrow$	*BIG*
○	*red*	$\uparrow$	*little*

The ○/□-switch correlates strongly with the *red*/blue switch

$\Rightarrow$these two genes must be close in the cell

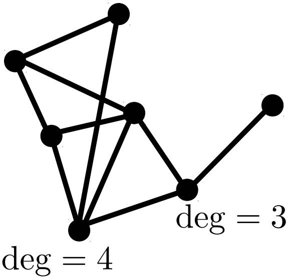
the corresponding metric space is 1-dimensional.

1944	Avery-MacLeod-McCarty
1952	Hershey-Chase
1953	Watson-Crick
$\sim$ 1959	Nirenberg, ...

DNA is a double helix carrying the genetic information

Quite often, in sciences, and in computer science alike, you want to infer properties of an object you cannot observe directly.

Here, we want to "observe" properties and parameters of graphs.



Property: YES/NO
planarity, containing a $\triangle$

Parameter: real number
chromatic number, no. of $\triangle$'s

What is the average number of "friends" in the Facebook graph?

(Graph) Parameter estimation

Setting:
Alice: Holds a (large) graph G.
Bob: Wants to learn a property/parameter $f(G)$. Wants to use as few queries as possible. (and no other computational restrictions)

(Graph) Parameter estimation

Setting:
Alice: Holds a (large) graph G.
Bob: Wants to learn a property/parameter $f(G)$. Wants to use as few queries as possible. (and no other computational restrictions)

A 1-query solution

Bob: Tell me $f(G)$.

(Graph) Parameter estimation

Setting:
Alice: Holds a (large) graph G.
Bob: Wants to learn a property/parameter $f(G)$. Wants to use as few queries as possible. (and no other computational restrictions)

A 1-query solution

Bob: Tell me $f(G)$.

So, to turn this into a non-trivial problem, we only allow queries of the type: *Is* $ij \in E(G)$*?*

(Graph) Parameter estimation

Setting:
Alice: Holds a (large) graph G.
Bob: Wants to learn a property/parameter $f(G)$. Wants to use as few queries as possible. (and no other computational restrictions)

A 1-query solution

Bob: Tell me $f(G)$.

So, to turn this into a non-trivial problem, we only allow queries of the type: *Is* $ij \in E(G)$*?*

Typically, it is impossible to determine $f(G)$ before learning the entire graph G (at least in the worst case). Rather, we want to get a **high-confidence** $(= 1 - \epsilon)$ **estimate on** $f(G)$ **using few** $(= K(\epsilon))$ **randomized queries**.

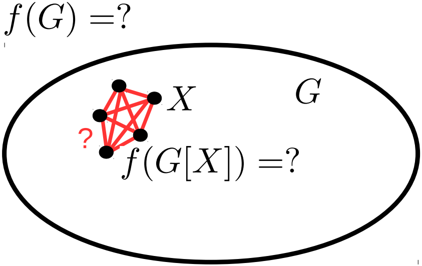
Parameter estimation in dense graphs

our universe: all graphs. . . $\mathcal{G}$

A parameter $f : \mathcal{G} \to \mathbb{R}$ is estimable if for every $\epsilon > 0$ there exists a number $K = K(\epsilon)$ such that

$$\mathbb{P}\big[f(G) - f(G[X])| > \epsilon\big] < \epsilon ,$$

where $X \subset V(G)$ is a random K-set.



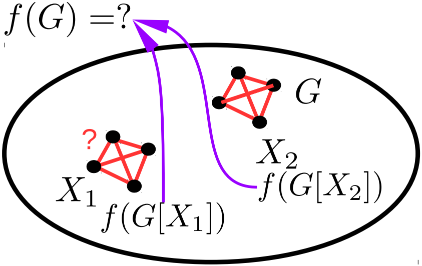
Parameter estimation in dense graphs

our universe: all graphs… $\mathcal{G}$

A parameter $f : \mathcal{G} \to \mathbb{R}$ is estimable if for every $\epsilon > 0$ there exists a number $K = K(\epsilon)$ such that

$$\mathbb{P}\big[f(G) - f(G[X])| > \epsilon\big] < \epsilon\,,$$

where $X \subset V(G)$ is a random K-set.



Parameter estimation in dense graphs

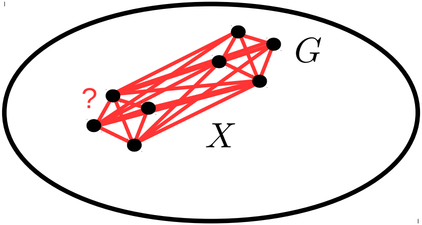
our universe: all graphs... $\mathcal{G}$

A parameter $f : \mathcal{G} \to \mathbb{R}$ is estimable if for every $\epsilon > 0$ there exists a number $K = K(\epsilon)$ such that

$$\mathbb{P}\big[f(G) - f(G[X])| > \epsilon\big] < \epsilon \, ,$$

where $X \subset V(G)$ is a random K-set.

$f(G) =?$



Parameter estimation in dense graphs

our universe: all graphs. . . $\mathcal{G}$

A parameter $f : \mathcal{G} \to \mathbb{R}$ is estimable if for every $\epsilon > 0$ there exists a number $K = K(\epsilon)$ such that

$$\mathbb{P}\big[f(G) - f(G[X])| > \epsilon\big] < \epsilon \,,$$

where $X \subset V(G)$ is a random K-set.

e.g. $f =$ triangle density $= \# \triangle / n^3$ is estimable

Parameter estimation in dense graphs

our universe: all graphs... $\mathcal{G}$
A parameter $f : \mathcal{G} \to \mathbb{R}$ is estimable if for every $\epsilon > 0$ there exists a number $K = K(\epsilon)$ such that

$$\mathbb{P}\big[f(G) - f(G[X])| > \epsilon\big] < \epsilon \,,$$

where $X \subset V(G)$ is a random K-set.

e.g. f = triangle density $= \# \triangle / n^3$ is estimable

Why **dense**?

Parameter estimation in dense graphs

our universe: all graphs. . . $\mathcal{G}$

A parameter $f : \mathcal{G} \to \mathbb{R}$ is estimable if for every $\epsilon > 0$ there exists a number $K = K(\epsilon)$ such that

$$\mathbb{P}\big[f(G) - f(G[X])| > \epsilon\big] < \epsilon \ ,$$

where $X \subset V(G)$ is a random K-set.

e.g. f = triangle density $= \# \triangle / n^3$ is estimable

Why **dense**? Recall: $e(G) \leq \binom{n}{2} \approx n^2/2$.

Observe that an estimable parameter cannot change substantially after an $o(n^2)$ edge-perturbation of G.

In particular, $f(G) \approx f(\emptyset)$, whenever $e(G) = o(n^2)$.

No information about trees, planar graphs, . . .

Limits of dense graph sequences

Lovász, Szegedy *JCTB'06* (Fulkerson Prize'12)
Borgs, Chayes, Lovász, Sós, Vesztergombi *STOC'06*
Borgs, Chayes, Lovász, Sós, Vesztergombi *Adv.Math.'06*
Borgs, Chayes, Lovász, Sós, Vesztergombi *Ann.Math.'12*

Limits of dense graph sequences

Lovász, Szegedy *JCTB'06* (Fulkerson Prize'12)
Borgs, Chayes, Lovász, Sós, Vesztergombi *STOC'06*
Borgs, Chayes, Lovász, Sós, Vesztergombi *Adv.Math.'06*
Borgs, Chayes, Lovász, Sós, Vesztergombi *Ann.Math.'12*

idea: convergence notion for sequences of finite graphs
compactification of the space of finite graphs $\Rightarrow$
... *graphons* symmetric Lebesgue-m. functions $\Omega^2 \to [0, 1]$

Why? same story as with $\mathbb{Q}$ vs $\mathbb{R}$: only the latter allows
reasonable e.g. variational and integral calculus
for example $argmin(x^3 - 2x)$

Limits of dense graph sequences

Lovász, Szegedy *JCTB'06* (Fulkerson Prize'12)
Borgs, Chayes, Lovász, Sós, Vesztergombi *STOC'06*
Borgs, Chayes, Lovász, Sós, Vesztergombi *Adv.Math.'06*
Borgs, Chayes, Lovász, Sós, Vesztergombi *Ann.Math.'12*

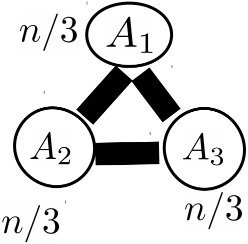
idea: convergence notion for sequences of finite graphs
compactification of the space of finite graphs $\Rightarrow$
... *graphons* symmetric Lebesgue-m. functions $\Omega^2 \to [0,1]$

Why? same story as with $\mathbb{Q}$ vs $\mathbb{R}$: only the latter allows
reasonable e.g. variational and integral calculus
for example $argmin(x^3 - 2x)$
mathematical framework for parameter estimation and property testing

Limits of dense graph sequences: an abstract approach

F is a "fixed graph" of order k, G is "large" of order n
We define **subgraph density** $t(F, G)$:

$$t(F,G) := \frac{\#\text{ copies of } F \text{ in } G}{\binom{n}{k}} = \mathbb{P}\big[G[\text{random } k\text{-set}] \cong F\big]$$



$k=2:\quad t(\;, G) = \frac{1}{3} \qquad t(\;, G) = \frac{2}{3}$

$k=3:\quad t(\;, G) = \frac{1}{9} \qquad t(\;, G) = 0$

$\qquad\quad\;\; t(\;, G) = \frac{2}{3} \qquad t(\;, G) = \frac{2}{9}$

$k=4:\quad t(\;, G) = 0 \qquad \dots$

Limits of dense graph sequences: an abstract approach

F is a "fixed graph" of order k, G is "large" of order n
We define **subgraph density** $t(F, G)$:

$$t(F,G) := \frac{\#\text{ copies of } F \text{ in } G}{\binom{n}{k}} = \mathbb{P}\big[G[\text{random } k\text{-set}] \cong F\big]$$

A sequence of graphs $G_1, G_2, \ldots$ **converges** if for each F, the sequence $t(F, G_1), t(F, G_2), \ldots$ converges.

We get a **limit object** Ψ, $t(F, \Psi) = \lim_n t(F, G_n)$.

Limits of dense graph sequences: an abstract approach

F is a "fixed graph" of order k, G is "large" of order n
We define **subgraph density** $t(F, G)$:

$$t(F,G) := \frac{\#\text{ copies of } F \text{ in } G}{\binom{n}{k}} = \mathbb{P}\big[G[\text{random } k\text{-set}] \cong F\big]$$

A sequence of graphs $G_1, G_2, \ldots$ **converges** if for each F, the sequence $t(F, G_1), t(F, G_2), \ldots$ converges.

We get a **limit object** Ψ, $t(F, \Psi) = \lim_n t(F, G_n)$.

Topology on the limit space: $\mathrm{dist}(\Psi_1, \Psi_2) \leq 1/k$, if the total variation distance of $\{t(F, \Psi_1)\}_{v(F)=k}$ and $\{t(F, \Psi_2)\}_{v(F)=k}$ is at most $1/k$.
In particular, we can measure distance between finite graphs.

Limits of dense graph sequences: an abstract approach

F is a "fixed graph" of order k, G is "large" of order n
We define **subgraph density** $t(F, G)$:

$$t(F,G) := \frac{\#\text{ copies of } F \text{ in } G}{\binom{n}{k}} = \mathbb{P}\big[G[\text{random } k\text{-set}] \cong F\big]$$

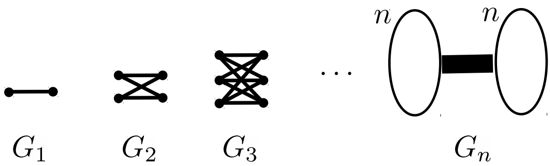
A sequence of graphs $G_1, G_2, \ldots$ **converges** if for each F, the sequence $t(F, G_1), t(F, G_2), \ldots$ converges.

We get a **limit object** Ψ, $t(F, \Psi) = \lim_n t(F, G_n)$.

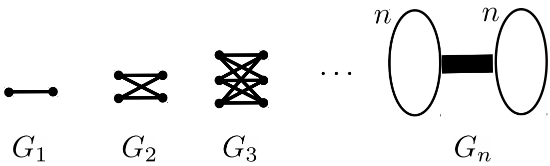
Topology on the limit space: $\text{dist}(\Psi_1, \Psi_2) \leq 1/k$, if the total variation distance of $\{t(F, \Psi_1)\}_{v(F)=k}$ and $\{t(F, \Psi_2)\}_{v(F)=k}$ is at most $1/k$.
In particular, we can measure distance between finite graphs.

The key connection: $f : \mathcal{G} \to \mathbb{R}$ is estimable iff it is continuous.

Graphons

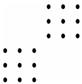


Graphons

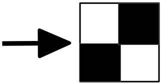


Represent these graphs by their adjacency matrices:

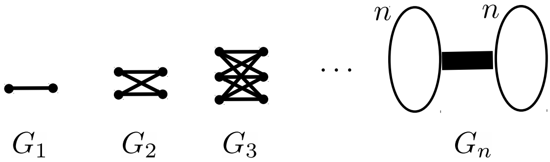
$$\begin{matrix}0 & 1\\ 1 & 0\end{matrix} \qquad \begin{matrix}0 & 0 & 1 & 1\\ 0 & 0 & 1 & 1\\ 1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\end{matrix}$$



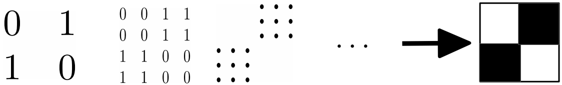
$\cdots$



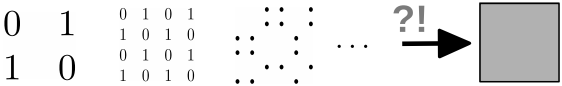
Graphons



Represent these graphs by their adjacency matrices:



. . . works if you do things the right way. But, . . .



In general Szemerédi's Regularity Lemma can be used to determine "the right way" of ordering the vertices.

Dense model

complete picture:

characterization of testable graph properties and estimable parameters either

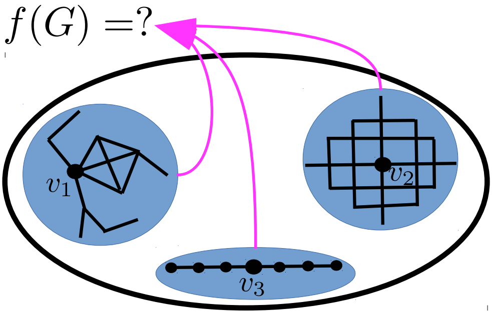
- in the language of the Szemerédi Regularity lemma ($\subseteq$Alon–Fischer–Newman–Shapira'03–'06, ...), and
- in the language of graph limits.

Parameter estimation in bounded degree graphs

our universe: graphs of degrees bounded by a constant D . . . $\mathcal{G}_D$
A parameter $f : \mathcal{G}_D \to \mathbb{R}$ is estimable if for each $\epsilon > 0$ there exists a number $K = K(\epsilon)$ and a function g such that

$$\mathbb{P}\left[|f(G) - g(B_1, B_2, \ldots, B_K)| > \epsilon\right] < \epsilon \ ,$$

where $B_1, \ldots, B_K$ are balls of radius K around K randomly selected vertices of G.

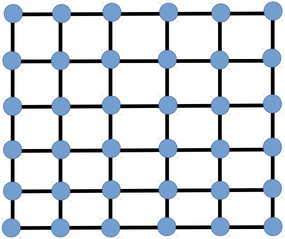


Limits of sparse graph sequences

$G_1, G_2, G_3, \ldots \in \mathcal{G}_D$.
Goal: convergence notion.

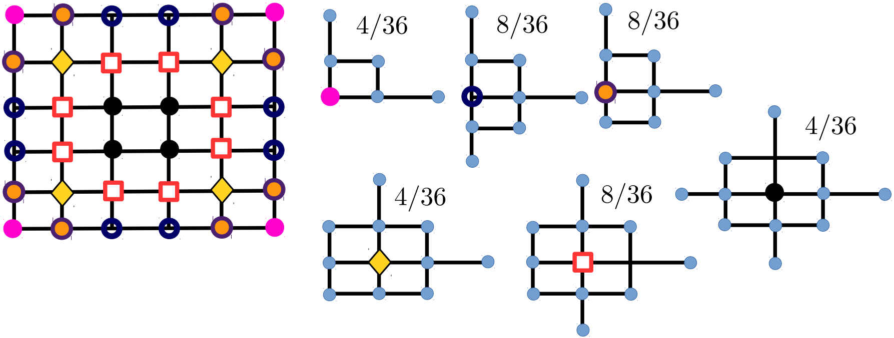
Limits of sparse graph sequences

$G_1, G_2, G_3, \ldots \in \mathcal{G}_D$.
Goal: convergence notion.
$\rho_r(G)$ =distribution on rooted r-balls around a randomly selected root of G. (example $r = 2$)



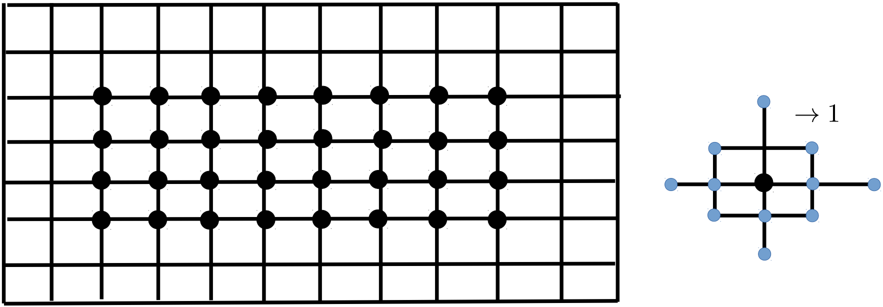
Limits of sparse graph sequences

$G_1, G_2, G_3, \ldots \in \mathcal{G}_D$.
Goal: convergence notion.
$\rho_r(G)$ =distribution on rooted r-balls around a randomly selected root of G. (example $r = 2$)



Limits of sparse graph sequences

$G_1, G_2, G_3, \ldots \in \mathcal{G}_D$.
Goal: convergence notion.
$\rho_r(G)$ =distribution on rooted r-balls around a randomly selected root of G. (example $r = 2$)



Limits of sparse graph sequences

$G_1, G_2, G_3, \ldots \in \mathcal{G}_D$.

Goal: convergence notion.

$\rho_r(G)$ =distribution on rooted r-balls around a randomly selected root of G. (example $r = 2$)

Definition: $G_1, G_2, G_3, \ldots$ is **convergent** if for each $r \in \mathbb{N}$, $\rho_r(G_1), \rho_r(G_2), \rho_r(G_3), \ldots$ converges (and converges to a probability distribution) **(Benjamini–Schramm'01)**

Limits of sparse graph sequences

$G_1, G_2, G_3, \ldots \in \mathcal{G}_D$.
Goal: convergence notion.
$\rho_r(G)$ =distribution on rooted r-balls around a randomly selected root of G. (example $r = 2$)
Definition: $G_1, G_2, G_3, \ldots$ is **convergent** if for each $r \in \mathbb{N}$, $\rho_r(G_1), \rho_r(G_2), \rho_r(G_3), \ldots$ converges (and converges to a probability distribution) **(Benjamini–Schramm'01)**

The key connection: $f : \mathcal{G}_D \to \mathbb{R}$ is estimable iff it is continuous.

Estimable parameters in the bounded-degree model

Negative example The independence ratio $\alpha(G)/n$ is NOT estimable.
($\alpha(G)$ =maximum size independent set. . . vertices induce no edge)

Estimable parameters in the bounded-degree model

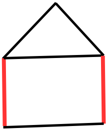
Negative example The independence ratio $\alpha(G)/n$ is NOT estimable.
($\alpha(G)$ =maximum size independent set. . . vertices induce no edge)

Thm [Nguyen, Onak, FOCS'08]: Matching ratio is estimable.

Estimable parameters in the bounded-degree model

Negative example The independence ratio $\alpha(G)/n$ is NOT estimable.
($\alpha(G)$ =maximum size independent set... vertices induce no edge)

Thm [Nguyen, Onak, FOCS'08]: Matching ratio is estimable.
(matching ratio=maximum matching$/n \in [0, \frac{1}{2}]$)



Estimable parameters in the bounded-degree model

Negative example The independence ratio $\alpha(G)/n$ is NOT estimable.
($\alpha(G)$ =maximum size independent set...vertices induce no edge)

Thm [Nguyen, Onak, FOCS'08]: Matching ratio is estimable.
Proof: Construct a suitable estimator, and prove that with high probability it gives a good estimate for the matching ratio

Estimable parameters in the bounded-degree model

Negative example The independence ratio $\alpha(G)/n$ is NOT estimable.
($\alpha(G)$ =maximum size independent set. . . vertices induce no edge)

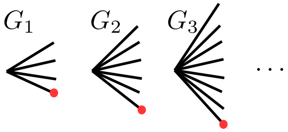
Thm [Nguyen, Onak, FOCS'08]: Matching ratio is estimable.
Proof: Construct a suitable estimator, and prove that with high probability it gives a good estimate for the matching ratio
Proof [Elek–Lippner]: (Borel oracles method)
Argue that there exists a "Borel matching" on the limit space.
Show how to make use of this structure to make estimates about matching ratio of finite graphs.
In particular, this does not give any construction of an algorithm!

Bounded degrees

Why did we have to have all degrees$\leq D$?

Bounded degrees

Why did we have to have all degrees $\leq D$?

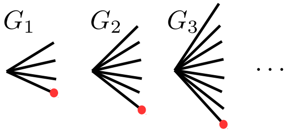


Random rooted 2-balls $G_1, G_2, G_3, \ldots$ have a weak limit, but a trivial one (total mass$=0$).

Maximum degree$\leq D \Rightarrow$ finitely many r-balls
$\Rightarrow$ measure cannot "escape to infinity"

Bounded degrees

Why did we have to have all degrees$\leq D$?



Random rooted 2-balls $G_1, G_2, G_3, \dots$ have a weak limit, but a trivial one (total mass$=0$).

Maximum degree$\leq D \Rightarrow$ finitely many r-balls

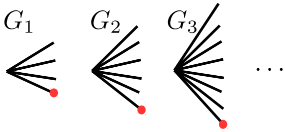
$\Rightarrow$ measure cannot "escape to infinity"

A sequence of probability measures $\mu_1, \mu_2, \dots$ on $\mathcal{X}$ is **tight** if for every $\epsilon > 0$ there exists a **finite** $K \subset \mathcal{X}$ such that $\mu_n(K) \geq 1 - \epsilon$ for all n.

Lyons'07: The concept of Benjamini–Schramm limit can be extended to sequences $G_1, G_2, \dots$ where for each $r \in \mathbb{N}$, the sequence $\rho_r(G_1), \rho_r(G_2), \dots$ is tight. AND NOT FURTHER

Bounded degrees

Why did we have to have all degrees$\leq D$?



Random rooted 2-balls $G_1, G_2, G_3, \ldots$ have a weak limit, but a trivial one (total mass$=$0).

Maximum degree$\leq D \Rightarrow$ finitely many r-balls

$\Rightarrow$ measure cannot "escape to infinity"

A sequence of probability measures $\mu_1, \mu_2, \ldots$ on $\mathcal{X}$ is **tight** if for every $\epsilon > 0$ there exists a **finite** $K \subset \mathcal{X}$ such that $\mu_n(K) \geq 1 - \epsilon$ for all n.

Lyons'07: The concept of Benjamini–Schramm limit can be extended to sequences $G_1, G_2, \ldots$ where for each $r \in \mathbb{N}$, the sequence $\rho_r(G_1), \rho_r(G_2), \ldots$ is tight. AND NOT FURTHER

with Lukasz Grabowski and Oleg Pikhurko, 2014+

the limit space and the soft arguments in the theory of bounded-degree graph limits make sense even for tight graph sequences

New graph classes for which the graph limit used not to be applicable:

- Erdős–Rényi $\mathbb{G}_{n,C/n}$,
- random planar graphs,...

no surprises yet.