Recent Progress on Random Graph Matching Problems

Zhangsong Li

School of Mathematical Sciences
Peking University

May 20, 2023

Based on a joint work with J.Ding

Application 1: Network de-anonymization

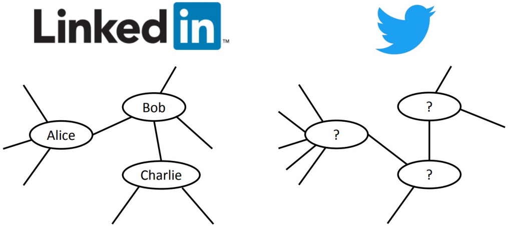


Figure 1: Picture courtesy of R.Srikant

Application 1: Network de-anonymization

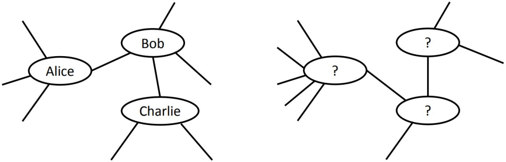


Figure 1: Picture courtesy of R.Srikant

- Successfully de-anonymize Netflix by matching it to IMDB. [Narayanan-Shmatikov '08]

Application 1: Network de-anonymization

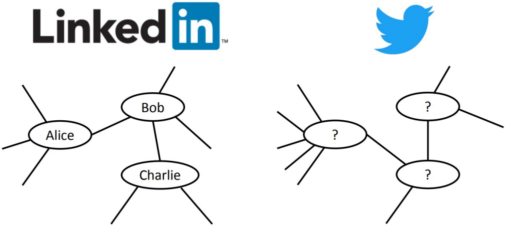


Figure 1: Picture courtesy of R.Srikant

- Successfully de-anonymize Netflix by matching it to IMDB. [Narayanan-Shmatikov '08]
- Correctly identified 30.8% of node mappings between Twitter and Flickr. [Narayanan-Shmatikov '09]

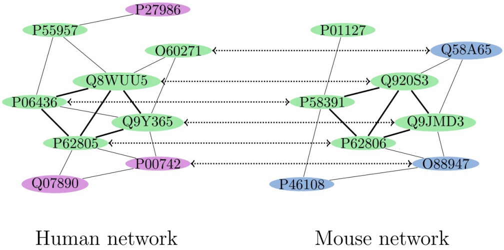
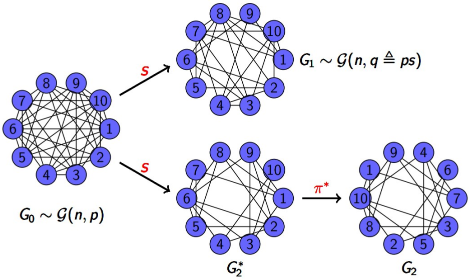


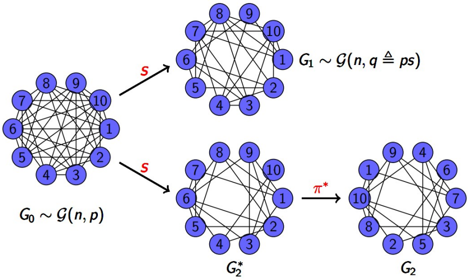
Figure 2: [Kazemi-Hassani-Grossglauser-Modarres '16]

- Ontology: Discover proteins with similar functions across different species based interaction network topology.

An idealized model: correlated Erdős-Rényi graphs

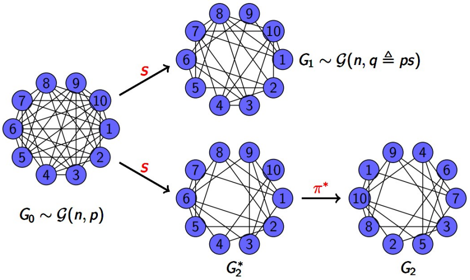


An idealized model: correlated Erdős-Rényi graphs



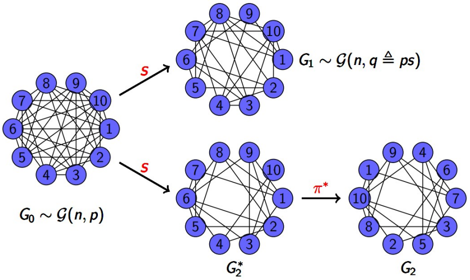
- There is no structure in randomness: there is an edge between each pair of vertices with probability p independently.

An idealized model: correlated Erdős-Rényi graphs



- There is no structure in randomness: there is an edge between each pair of vertices with probability p independently.
- Advantage: simple probabilistic model; suitable playground for developing mathematical theory.

An idealized model: correlated Erdős-Rényi graphs



- There is no structure in randomness: there is an edge between each pair of vertices with probability p independently.
- Advantage: simple probabilistic model; suitable playground for developing mathematical theory.
- Disadvantage: almost all realistic networks are not Erdős-Rényi.

Information-computation phase transition

Information-computation phase transition

- Intuitively, the matching π^* that maximize the common edge between two graphs (i.e. the MLE) should be the most effective estimator for recovering the latent matching π.

Information-computation phase transition

- Intuitively, the matching π^* that maximize the common edge between two graphs (i.e. the MLE) should be the most effective estimator for recovering the latent matching π.
- Reduced the problem into the network alignment problem of correlated random graphs.

Information-computation phase transition

- Intuitively, the matching π^* that maximize the common edge between two graphs (i.e. the MLE) should be the most effective estimator for recovering the latent matching π.
- Reduced the problem into the network alignment problem of correlated random graphs.
- Unfortunately, the classical graph alignment problem is a NP-hard optimization problem, so we must seek help from randomness.

Quick overview on information threshold

- **Three thresholds**: detection, partial recovery, exact recovery;

Quick overview on information threshold

- **Three thresholds**: detection, partial recovery, exact recovery;
- [Wu-Xu-Yu'20,21]: progress based analyzing MLE (see also Ganassali-Massoulié-Lelarge for $p \approx 1/n$).

Quick overview on information threshold

- **Three thresholds**: detection, partial recovery, exact recovery;
- [Wu-Xu-Yu'20,21]: progress based analyzing MLE (see also Ganassali-Massoulié-Lelarge for $p \approx 1/n$).
- Results: determine the exact information threshold for exact recovery; Determine the information threshold for partial-recovery and detection in the dense region ($p = n^{o(1)}$) exactly and in the non-dense region ($p = n^{-\alpha+o(1)}$ where $0 < \alpha < 1$) up to constants.

Quick overview on information threshold

- **Three thresholds**: detection, partial recovery, exact recovery;
- [Wu-Xu-Yu'20,21]: progress based analyzing MLE (see also Ganassali-Massoulié-Lelarge for $p \approx 1/n$).
- Results: determine the exact information threshold for exact recovery; Determine the information threshold for partial-recovery and detection in the dense region ($p = n^{o(1)}$) exactly and in the non-dense region ($p = n^{-\alpha+o(1)}$ where $0 < \alpha < 1$) up to constants.
- [Ding-Du'22+a,22+b]: determine the exact information threshold for detection and partial-recovery in the non-dense region via a modified statistics based on densest subgraphs.

Matching algorithms for correlated graphs (up to 2021)

Matching algorithms for correlated graphs (up to 2021)

- “Signature” based algorithm: for each vertex, compute a “signature” and match pairs of vertices with similar signatures. Desired properties for signature: **informative, computable, tractable**, generalizable.

Matching algorithms for correlated graphs (up to 2021)

- "Signature" based algorithm: for each vertex, compute a "signature" and match pairs of vertices with similar signatures. Desired properties for signature: **informative, computable, tractable**, generalizable.
- [Dai-Cullina-Kiyavash-Grossglauser'18] [Barak-Chou-Lei-Schramm-Sheng'19] [Ding-Mao-Wu-Xu'21].

Matching algorithms for correlated graphs (up to 2021)

- "Signature" based algorithm: for each vertex, compute a "signature" and match pairs of vertices with similar signatures. Desired properties for signature: **informative, computable, tractable**, generalizable.
- [Dai-Cullina-Kiyavash-Grossglauser'18] [Barak-Chou-Lei-Schramm-Sheng'19] [Ding-Mao-Wu-Xu'21].
- "Optimization relaxation" based algorithm:
  - [Fan-Mao-Wu-Xu'19+]. Original optimization problem is hard to solve, but feasible if enlarge the space of potential solutions (e.g. to a convex space).

Matching algorithms for correlated graphs (up to 2021)

- "Signature" based algorithm: for each vertex, compute a "signature" and match pairs of vertices with similar signatures. Desired properties for signature: **informative, computable, tractable**, generalizable.
- [Dai-Cullina-Kiyavash-Grossglauser'18] [Barak-Chou-Lei-Schramm-Sheng'19] [Ding-Mao-Wu-Xu'21].
- "Optimization relaxation" based algorithm:
  - [Fan-Mao-Wu-Xu'19+]. Original optimization problem is hard to solve, but feasible if enlarge the space of potential solutions (e.g. to a convex space).
- All the above algorithms either run in pseudo-polynomial time (i.e. $n^{O(\log n)}$) or succeeds only when the correlation approaches 1 (with rate polylog n).

Recent progress on matching algorithm

Recent progress on matching algorithm

- [Mao-Rudelson-Tikhomirov'21+]: poly-time algorithm based on partition trees, when p in particular region and correlation ≥ 0.99 (constant close to 1).

Recent progress on matching algorithm

- [Mao-Rudelson-Tikhomirov'21+]: poly-time algorithm based on partition trees, when p in particular region and correlation ≥ 0.99 (constant close to 1).
- [Ganassali-Massoulié-Lelarge'20+,22+]: poly-time partial matching algorithm for sparse graphs based on message passing, when correlation $\geq \sqrt{\text{Otter's constant}} \approx \sqrt{0.338}$.

Recent progress on matching algorithm

- [Mao-Rudelson-Tikhomirov'21+]: poly-time algorithm based on partition trees, when p in particular region and correlation ≥ 0.99 (constant close to 1).
- [Ganassali-Massoulié-Lelarge'20+,22+]: poly-time partial matching algorithm for sparse graphs based on message passing, when correlation $\geq \sqrt{\text{Otter's constant}} \approx \sqrt{0.338}$.
- [Mao-Wu-Xu-Yu'22+]: poly-time algorithm when correlation $> \sqrt{\text{Otter's constant}}$, based on a carefully curated family of rooted trees called chandeliers (substantially improving MRT21+, and covers much wider parameter regime).

Our contributions

Our contributions

- Based on previous works (especially MWXY22+, GML22+), you might guess the computation threshold for random graph matching is indeed given by the Otter's constant.

Our contributions

- Based on previous works (especially MWXY22+, GML22+), you might guess the computation threshold for random graph matching is indeed given by the Otter's constant.
- [Ding-L.'22+]: poly-time iterative algorithm for matching Gaussian matrices when correlation is non-vanishing.

Our contributions

- Based on previous works (especially MWXY22+, GML22+), you might guess the computation threshold for random graph matching is indeed given by the Otter's constant.
- [Ding-L.'22+]: poly-time iterative algorithm for matching Gaussian matrices when correlation is non-vanishing.
  - New feature: signal is stored in a vector where each coordinate is a pair of sets, and signal per coordinate decreases with iteration but compensated by increase on dimension.

Our contributions

- Based on previous works (especially MWXY22+, GML22+), you might guess the computation threshold for random graph matching is indeed given by the Otter's constant.
- [Ding-L.'22+]: poly-time iterative algorithm for matching Gaussian matrices when correlation is non-vanishing.
  - New feature: signal is stored in a vector where each coordinate is a pair of sets, and signal per coordinate decreases with iteration but compensated by increase on dimension.
  - Expected to be sharp, and should extend to graph matching (although with substantial challenge) assuming $np > n^{\alpha}$ for $\alpha > 0$.

Our contributions

- Based on previous works (especially MWXY22+, GML22+), you might guess the computation threshold for random graph matching is indeed given by the Otter's constant.
- [Ding-L.'22+]: poly-time iterative algorithm for matching Gaussian matrices when correlation is non-vanishing.
  - New feature: signal is stored in a vector where each coordinate is a pair of sets, and signal per coordinate decreases with iteration but compensated by increase on dimension.
  - Expected to be sharp, and should extend to graph matching (although with substantial challenge) assuming $np > n^{\alpha}$ for $\alpha > 0$.
  - Might shed lights on many matching problems too.

Our contributions

- Based on previous works (especially MWXY22+, GML22+), you might guess the computation threshold for random graph matching is indeed given by the Otter's constant.
- [Ding-L.'22+]: poly-time iterative algorithm for matching Gaussian matrices when correlation is non-vanishing.
  - New feature: signal is stored in a vector where each coordinate is a pair of sets, and signal per coordinate decreases with iteration but compensated by increase on dimension.
  - Expected to be sharp, and should extend to graph matching (although with substantial challenge) assuming $np > n^{\alpha}$ for $\alpha > 0$.
  - Might shed lights on many matching problems too.
- An ongoing work with J. Ding: A polynomial time iterative algorithm for random graph matching with non-vanishing correlation.

Perspectives and future directions

Perspectives and future directions

- A hub of theorists: combinatorics, probability, statistics, algorithms, complexity theory, optimization, etc.

Perspectives and future directions

- A hub of theorists: combinatorics, probability, statistics, algorithms, complexity theory, optimization, etc.
- A meeting point of theory and applications:

Perspectives and future directions

- A hub of theorists: combinatorics, probability, statistics, algorithms, complexity theory, optimization, etc.
- A meeting point of theory and applications:
  - Currently, most extensively studied models are idealistic. Even worse, many times algorithms and analysis are based on wrong model assumptions, e.g., local tree structure for social network model.

Perspectives and future directions

- A hub of theorists: combinatorics, probability, statistics, algorithms, complexity theory, optimization, etc.
- A meeting point of theory and applications:
  - Currently, most extensively studied models are idealistic. Even worse, many times algorithms and analysis are based on wrong model assumptions, e.g., local tree structure for social network model.
  - Major challenge 1: propose models with general applicability where theorists can say something meaningful.

Perspectives and future directions

- A hub of theorists: combinatorics, probability, statistics, algorithms, complexity theory, optimization, etc.
- A meeting point of theory and applications:
  - Currently, most extensively studied models are idealistic. Even worse, many times algorithms and analysis are based on wrong model assumptions, e.g., local tree structure for social network model.
  - Major challenge 1: propose models with general applicability where theorists can say something meaningful.
  - Major challenge 2: propose models for important scientific problems worth extensive theoretic study.

References

1. [BCL+19] B. Barak, C.-N. Chou, Z. Lei, T. Schramm, and Y. Sheng. (nearly) efficient algorithms for the graph matching problem on correlated random graphs. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.
2. [DCKG19] O. E. Dai, D. Cullina, N. Kiyavash, and M. Grossglauser. Analysis of a canonical labeling algorithm for the alignment of correlated erdos-rényi graphs. Proc. ACM Meas. Anal. Comput. Syst., 3(2), jun 2019.
3. [DD22+a] J. Ding and H. Du. Detection threshold for correlated erdos-renyi graphs via densest subgraph. IEEE Trans. Inform. Theory, online 2023.
4. [DD22+b] J. Ding and H. Du. Matching recovery threshold for correlated random graphs. Preprint, arXiv:2205.14650.
5. [DL22+] J. Ding and Z. Li. A polynomial time iterative algorithm for matching Gaussian matrices with non-vanishing correlation. Preprint, arXiv:2212.13677.
6. [DMWX21] J. Ding, Z. Ma, Y. Wu, and J. Xu. Efficient random graph matching via degree profiles. Probab. Theory Related Fields, 179(1-2):29–115, 2021.
7. [GM20] L. Ganassali and L. Massouli´e. From tree matching to sparse graph alignment. In J. Abernethy and S. Agarwal, editors, Proceedings of Thirty Third Conference on Learning Theory, volume 125 of Proceedings of Machine Learning Research, pages 1633–1665. PMLR, 09–12 Jul 2020.
8. [GMS22] L. Ganassali, L. Massouli´e, and G. Semerjian. Statistical limits of correlation detection in trees. arXiv:2209.13723.

References+

9. [MRT21+] C. Mao, M. Rudelson, and K. Tikhomirov. Exact matching of random graphs with constant correlation. Preprint, arXiv:2110.05000.
10. [MWXY22+] C. Mao, Y. Wu, J. Xu, and S. H. Yu. Random graph matching at Otter's threshold via counting chandeliers. Preprint, arXiv:2209.12313.
11. [MWXY21+] C. Mao, Y. Wu, J. Xu, and S. H. Yu. Testing network correlation efficiently via counting trees. Preprint, arXiv:2110.11816.
12. [NS08] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008), pages 111–125, 2008.
13. [NS09] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In 2009 30th IEEE Symposium on Security and Privacy, pages 173–187, 2009
14. [WXY21+] Y. Wu, J. Xu, and S. H. Yu. Settling the sharp reconstruction thresholds of random graph matching. Preprint, arXiv:2102.00082.

Thank you!