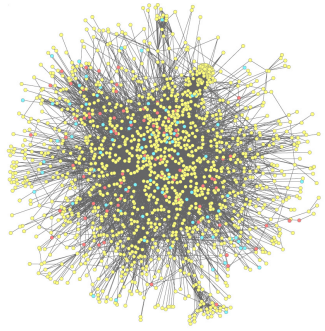
Inferring and Analyzing the Present and the Past of Networks from Limited Uncertain Data

Emre Sefer

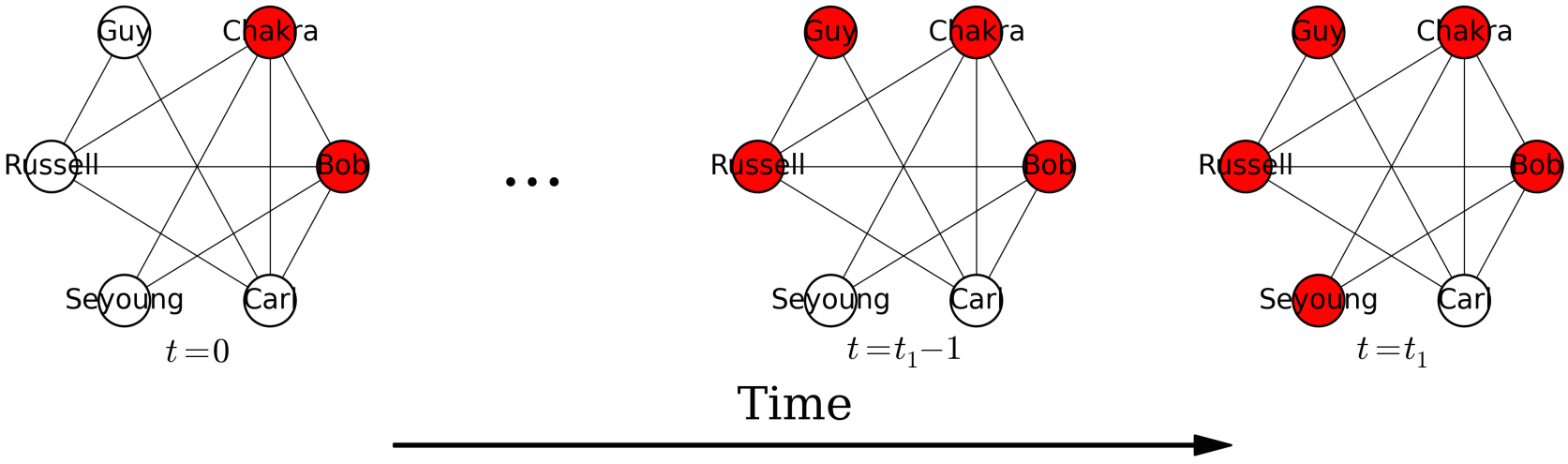
Thesis Committee
Carl Kingsford - CMU Chair
Russell Schwartz - CMU
Seyoung Kim - CMU
Chakra Chennubhotla - Pitt
Guy Blelloch - CMU

Thesis Summary



- We have interdependent dimensions over networks
  - Social network
    - diffusion over it $\rightarrow$ influenza diffusion
  - Protein-protein interaction network
    - node attributes $\rightarrow$ protein functions
  - Hi-C interaction network
    - data itself $\rightarrow$ ensemble of multiple matrices
- Thesis is motivated by problems over networks, between these dimensions

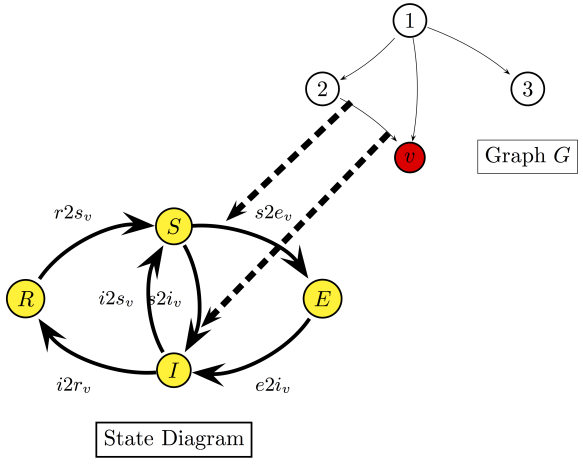
Networks & Diffusion



- Influenza diffusion between humans
- Idea diffusion over blog network
- Contaminant diffusion over water distribution network

SEIRS Dynamics

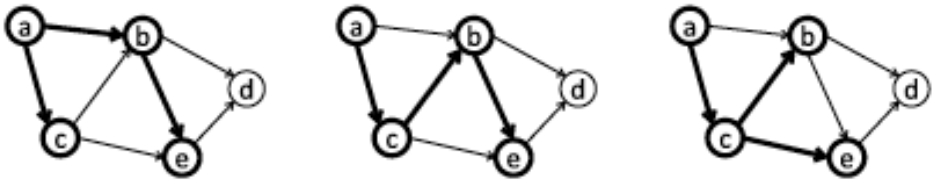
- SEIRS is a generalization of many diffusion models
  - SI, SIR, SIS, SEIR, SIRS, SEIRS
- Susceptible, exposed, infected, recovered

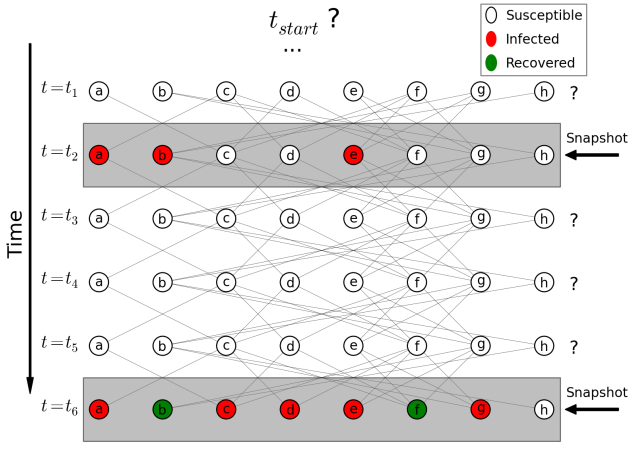


SEIRS State Transition Diagram

- Diffusion from single predecessor is enough to be affected

Two Diffusion Dynamics Problems





- Can we infer human-contact network from influenza diffusion data?[1]
- Can we reconstruct diffusion histories given only measurements at a few timepoints?[2]

[2] Sefer and Kingsford, "Convex Risk Minimization To Infer Networks From Probabilistic Diffusion Data At Multiple Scales".

[2] Sefer and Kingsford, "Diffusion Archaeology for Diffusion Progression History Reconstruction".

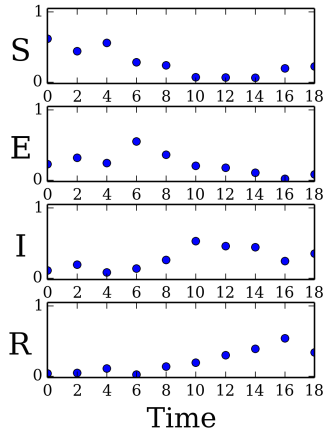
Why is Network Inference Important?

- Influenza transmission network
  - who infected whom

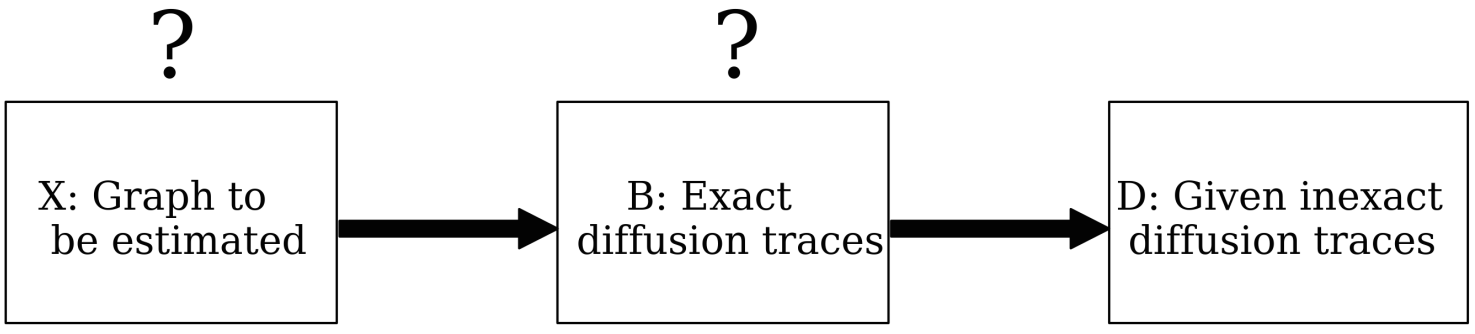
- Measuring contact network is harder in several scenarios
- Identify network motifs that are important in influenza diffusion

Network Reconstruction Problem

- Diffusion data can be inexact
  - Observed symptoms only partially reflect the true states

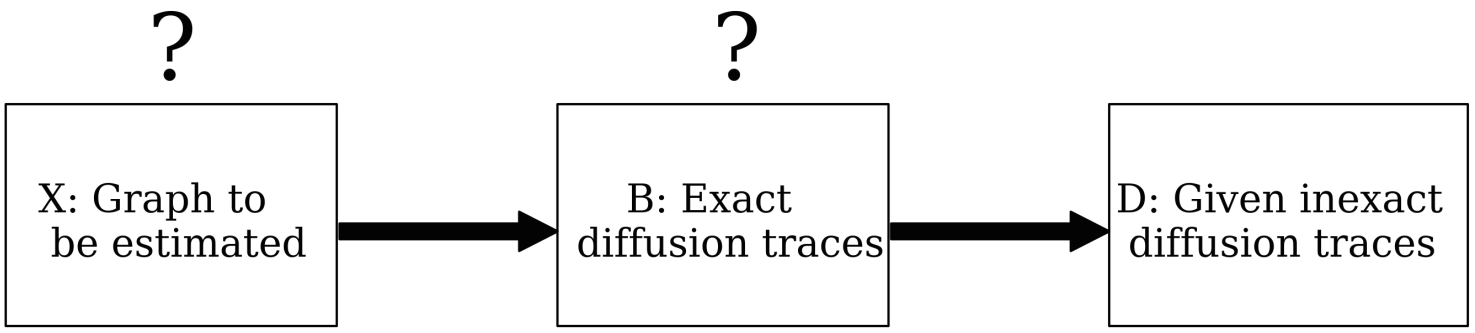


D



Network Reconstruction Problem

- *CORMIN*: Minimize the expected loss by convex optimization



$$R(X, D) = \mathbb{E}[X] = \sum_{B} \overbrace{L_B(X, B)}^{\text{Loss Function}} \overbrace{P(B|D)}^{\substack{\text{Pr.} \\ D \text{ is generated from } B}}$$

$$R(X, D) = \sum_{d \in D} R(X, d) = \sum_{d \in D} \sum_{b \in \mathcal{Q}(d)} L_b(X, b)\, P(b|d)$$

Human-Contact Network Prediction

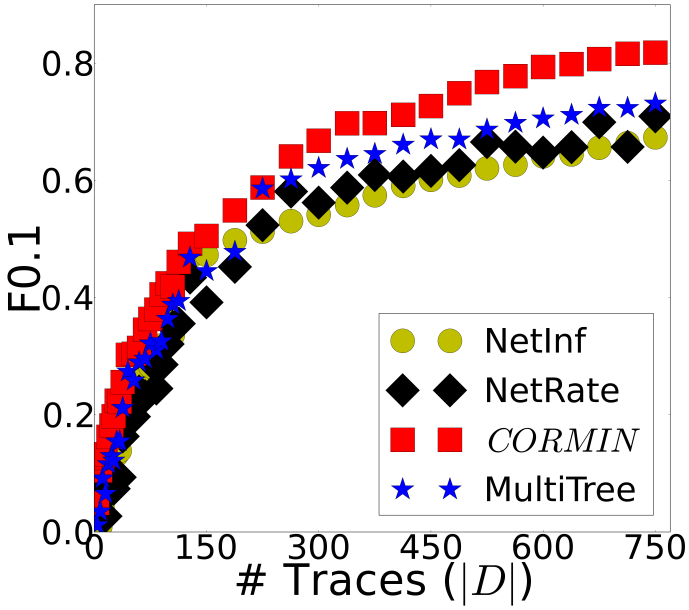
- Diffusion traces have realistic parameters:
  - p_{uv}: Weibull distribution with $\lambda = 9.5$, $k = 2.3$
  - p_v^{ei}, p_v^{ir}: Exponential distribution with $\lambda = 0.5$ and $\lambda = 0.2$
- Human-contact network has 750 nodes, 9777 edges[3]
  - Reconstruct edges from diffusion data

$$F0.1 = \frac{1.01\,\text{precision}\,\text{recall}}{0.01\text{precision} + \text{recall}}$$

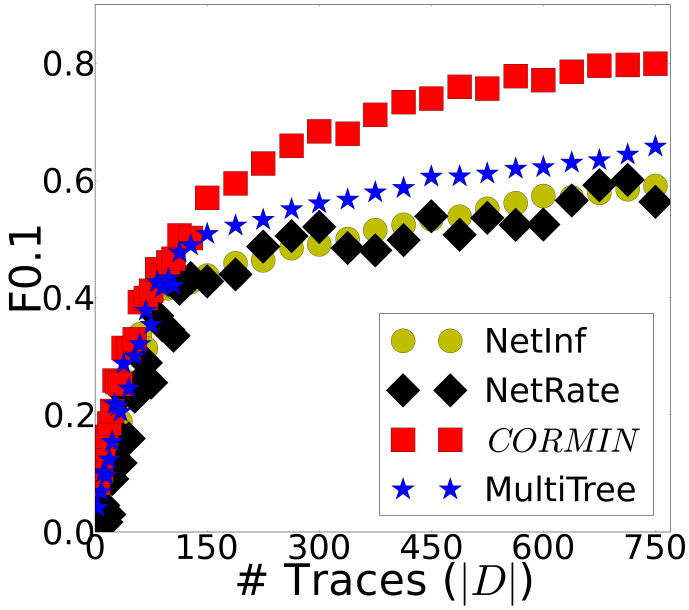
[3] Salathè et al., "A high-resolution Human Contact Network for Infectious Disease Transmission".

Accurate Prediction of Human-Contact Network

- We outperform the existing methods
  - We achieve F0.1 of 0.7 around 400 traces
- Similar performance for SEIR and $F1$



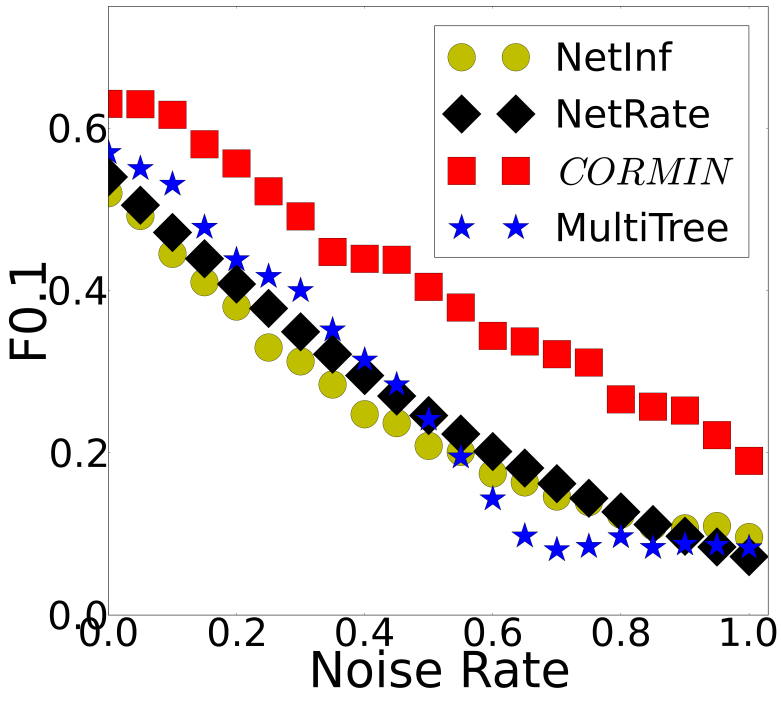
(a) SI perfect



(b) SIR perfect

Human-Contact Network Prediction

- We handle probabilistic diffusion data better than the existing methods



SI partial

Human-Contact Network Prediction

- Estimated network features are quite close to the true ones

	Estimated	Truth
Modularity [4]	0.67	0.73
Scale-free exponent	2.072	2.254
Assortativity	0.141	0.121
Avg. Clustering Coefficient	0.23	0.261
Diameter	10	8

[4] Newman, "Modularity and Community Structure in Networks".

Other Results

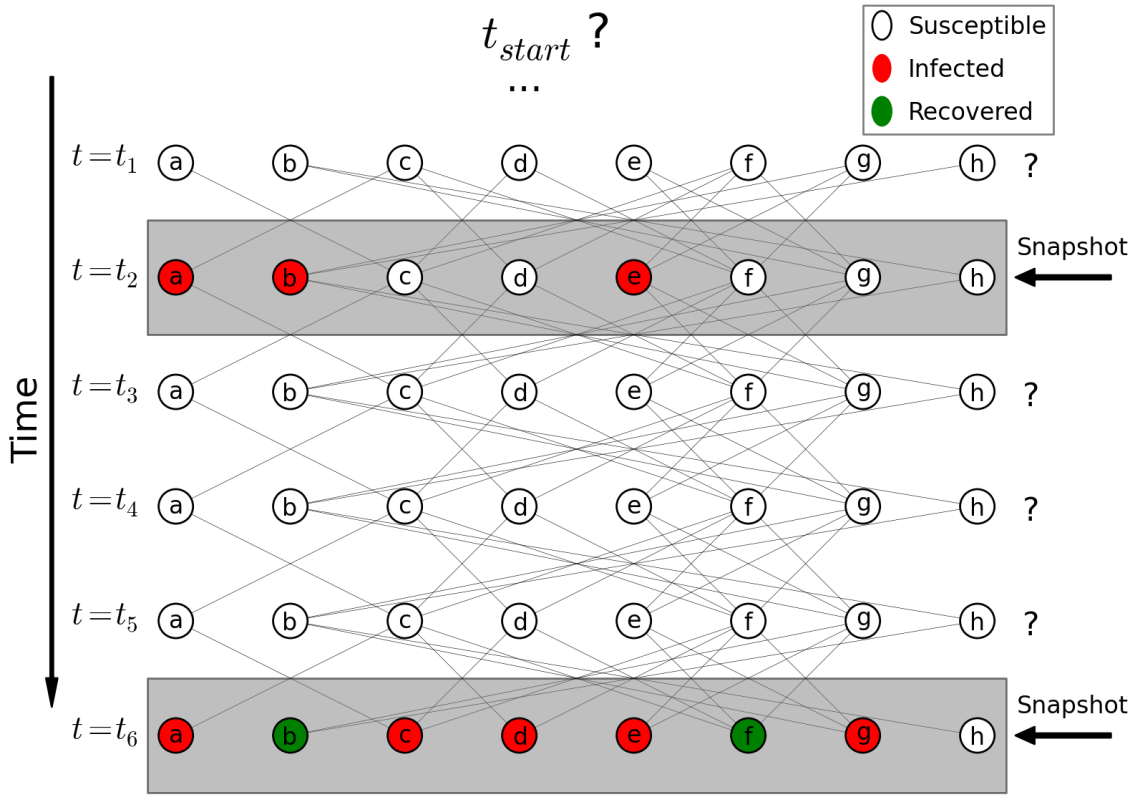
- At macroscale, we estimate diffusion rates between U.S. states by using Google Flu Trends data
  - We found $\rho = 0.32$ between estimated rates and true transportation rates
- Reasonable running time
  - Less than 5 minutes for 1000 node networks
- Inference for each node can be solved independently
  - Scalable parallelization of very large graph inference

Conclusion

- Improved recoverability of human-contact network
- We can better model both edge existence and nonexistence
- We formulate the problem at both micro- and macro- scales under SEIR
- Ability to handle noisy data

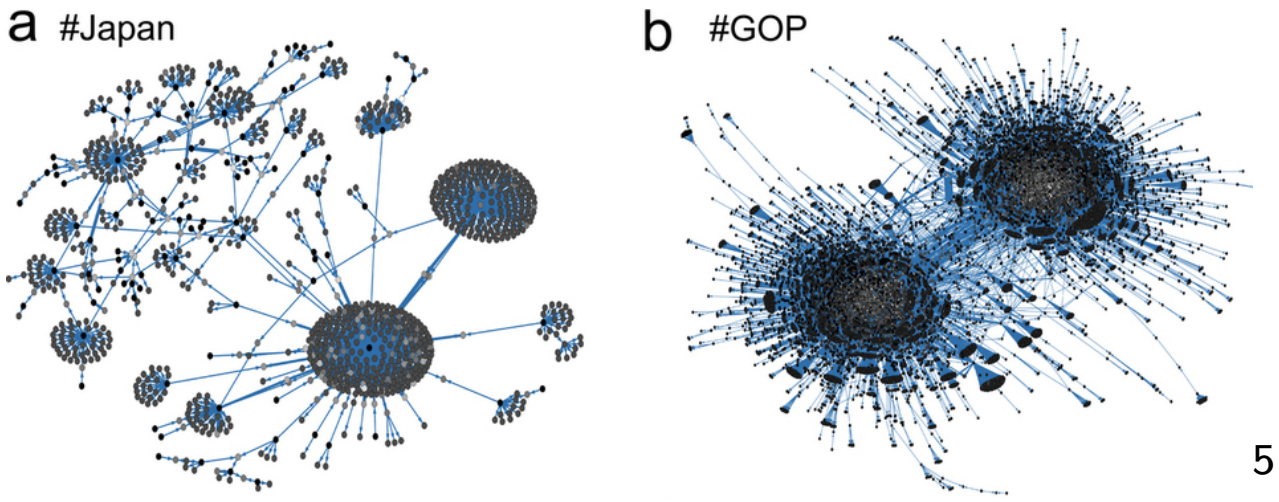
Reverse Question: Can we reconstruct diffusion histories?

- Can we reconstruct the diffusion histories over a graph from limited number of present-day information?



Why Important? Scenario 1

- Diffusion histories may not be completely available:
  - Physical limitations
  - Privacy
- We may not track the diffusion of memes between blogs
- We want to learn why a diffusion evolved to what we observed today?

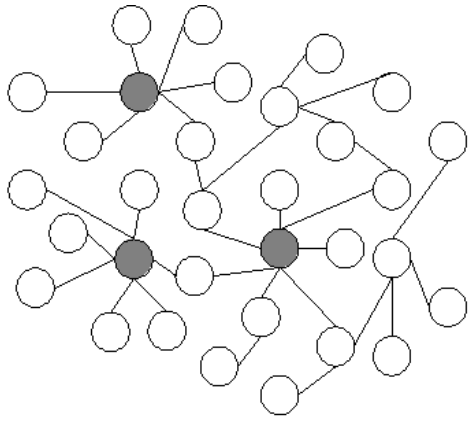


[5]

[5] Lillian Weng et al. "Competition among memes in a world with limited attention". In: *Scientific reports* 2 (2012).

Why Important? Scenario 2

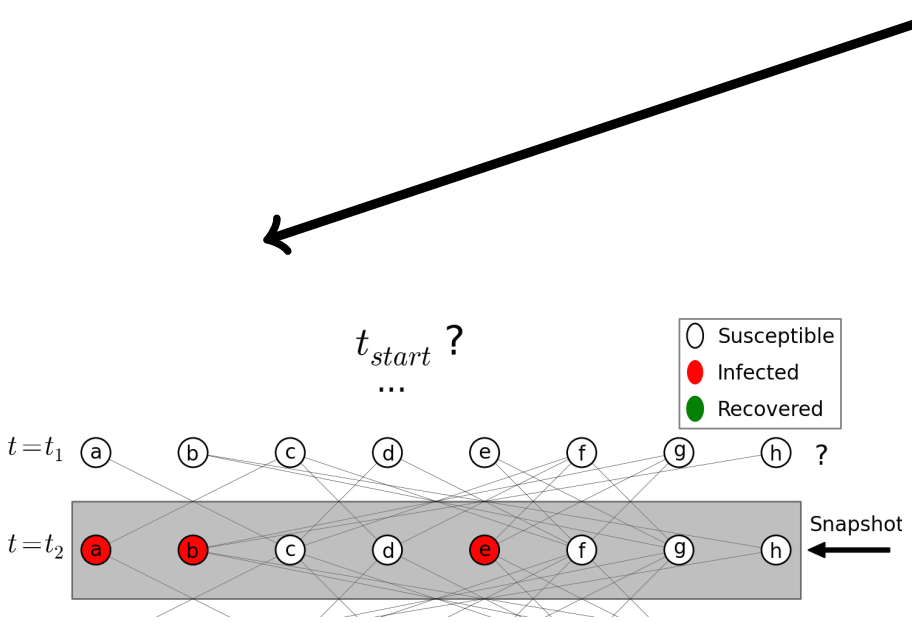
- Why did a computer virus go beyond the security and spread?
  - Which computers are the weakest ones?
- We may notice the diffusion only after a number of nodes become infected
  - Contaminant diffusion over water distribution network



Possible initial spreaders

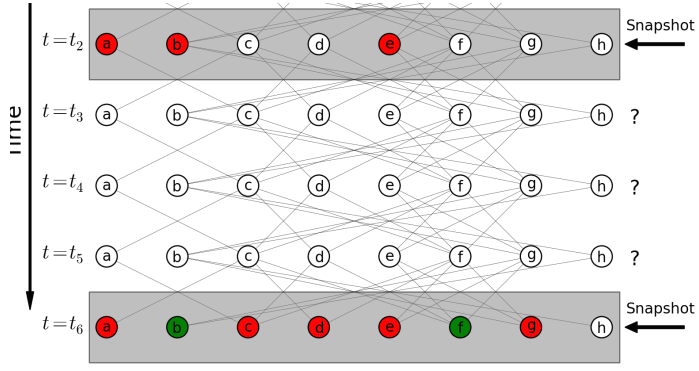
History Reconstruction (*DHR-sub*)

We solve two types of independent subproblems:



History Reconstruction Before the Earliest Snapshot

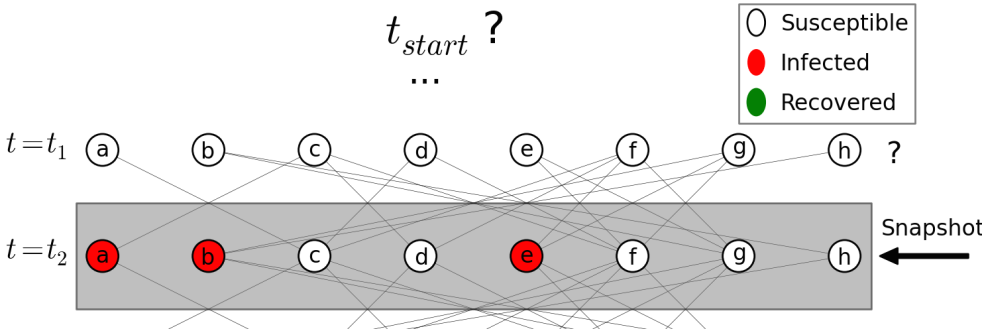
- Single problem



History Reconstruction Between Consecutive Snapshots

- Single problem between every consecutive time points

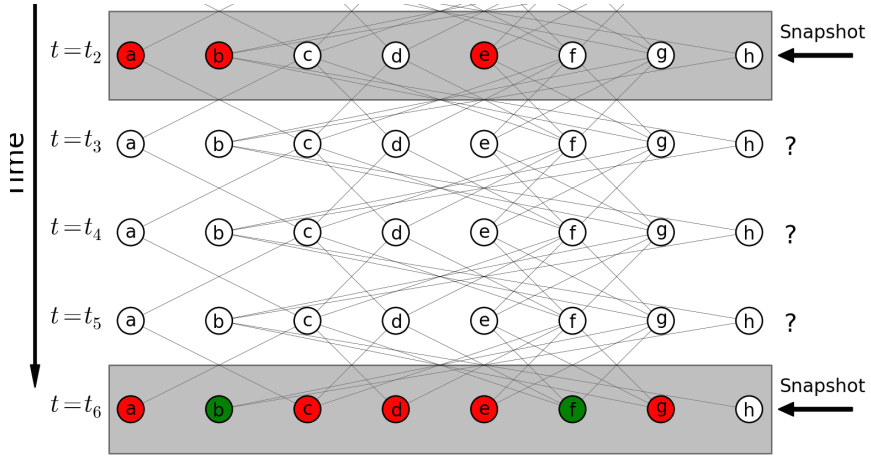
History Reconstruction Before the Earliest Snapshot



NP-Hard

- Initial diffusion time is unknown
- We greedily reconstruct the history at each previous time step by log-likelihood optimization
  - Unconstrained non-monotone submodular maximization for all SEIRS models except SIS
  - Normalized maximization version at each step can be approximated by $\frac{1}{3}$ via greedy method

History Reconstruction Between Consecutive Snapshots



NP-Hard

Given diffusion snapshots at time j and k;

- Problem is non-monotone submodular maximization under matroid base constraints
- We run modified greedy algorithm
  - Normalized maximization version can be approximated by $\frac{1}{6}$

Prize Collecting Relaxations (*DHR-pcdsvc*, *DHR-pcvc*)

- We need faster methods for large networks
- Replace log-likelihood by first-order Taylor relaxation
  - Problem becomes **Prize Collecting Dominating Set Vertex Cover** (*DHR-pcdsvc*)
  - We solve it by greedy method

Theorem

Prize Collecting Dominating Set Vertex Cover is NP-hard, and it can be approximated by $O(\log(|V^*|))$.

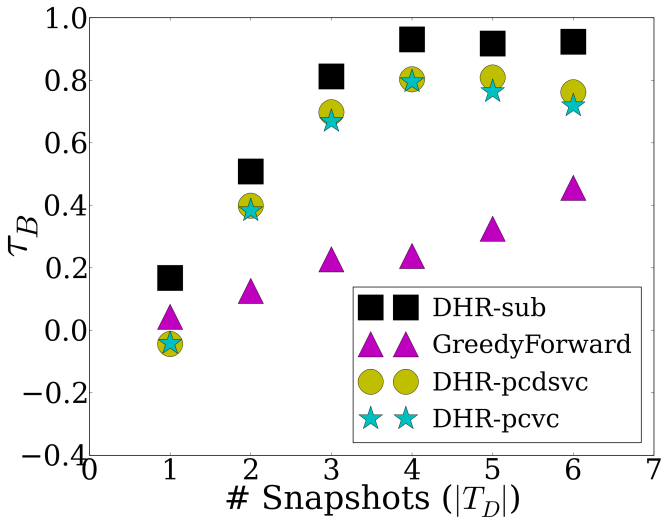
Experimental Evaluation

- Our methods are: ***DHR-sub***, ***DHR-pcdsvc***, ***DHR-pcvc***
- Competing methods: *Rumor*, *NetSleuth*, *Keffectors*
  - They cannot reconstruct the diffusion histories
- Greedy Baseline: *GreedyForward*

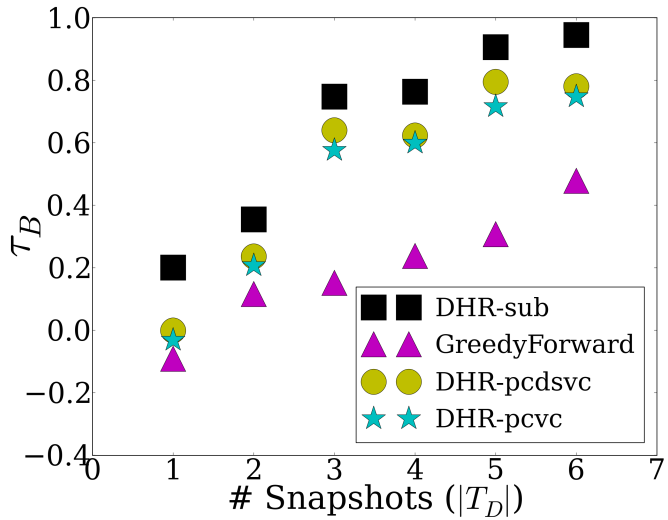
- Kendall Tau-b τ_B: measures the association between histories
  - -1 completely reverse, 1 perfect agreement
- Graph-based matching score $\overline{M}_G$: evaluates initial spreader prediction performance

Can we reconstruct the diffusion histories of memes?

- *Top-Blog* network: 5000 nodes[6]
- Performance is independent of the meme diffusion patterns
- Availability of even 2 or 3 snapshots increases the performance



a) *Fukushima*

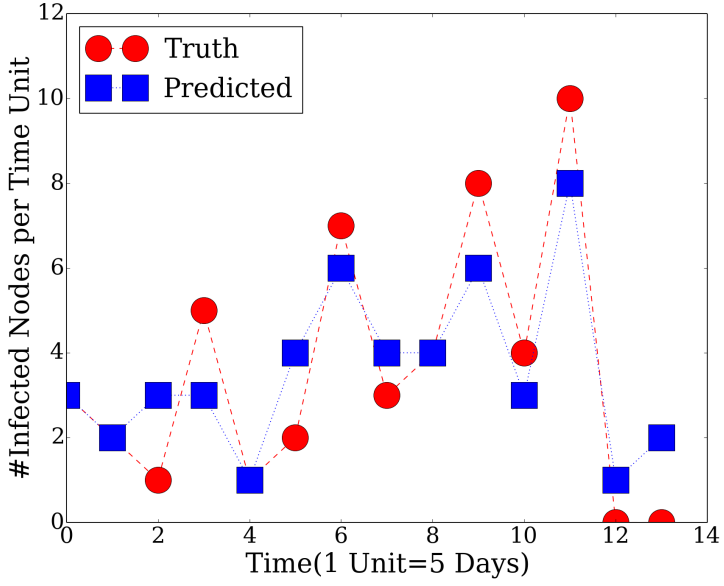


b) *Arab Spring*

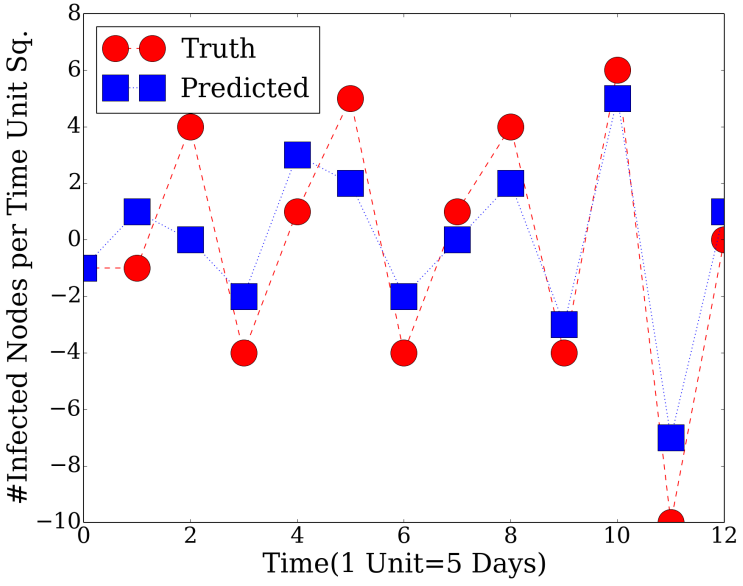
[6] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. "Structure and Dynamics of Information Pathways in Online Media". In: WSDM '13.

Can we predict temporal diffusion features?

- Diffusion of *Fukushima* over time from 3 snapshots
  - Reconstructed histories mimic closely their true speed and acceleration dynamics
  - Speed = #newly infected nodes per time
    Acceleration = Δ(Speed) per time



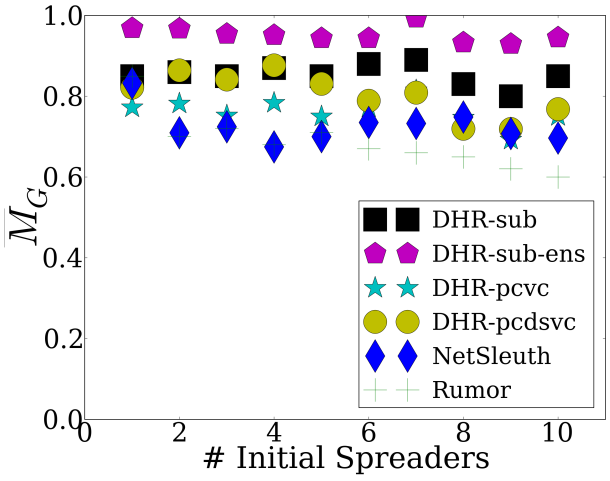
a) Speed



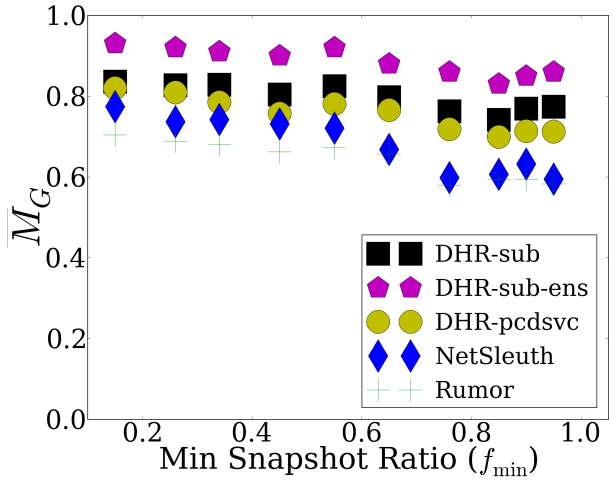
b) Acceleration

Can we identify the initial contamination locations?

- Generated contaminant diffusion by EPANET[7]
- Small water network (130 nodes), Large network (12527 nodes)[8]
  - Nodes are water demand-supply locations, edges are pipes
- Recoverability is robust to increasing diffusion length



a) *Water-sm*



b) *Water-big*

---

[7] Lewis A Rossman. "The EPANET ProgrammerâĂŹs Toolkit for Analysis of Water Distribution Systems". In: *ASCE 1999*.

[8] Avi Ostfeld et al. "The Battle of Water Sensor Networks (BWSN): A Design Challenge for Engineers and Algorithms". In: *Journal of Water Resources Planning and Management* (2008).

Other Results

- Our methods are scalable to networks up to 10000 nodes
  - We estimate meme histories in less than 5 minutes
- We perform quite accurate than baseline *GreedyForward* on different synthetic networks
- We can capture different shapes of speed and acceleration dynamics:
  - Uniform, bursty, repeating, etc

Conclusion

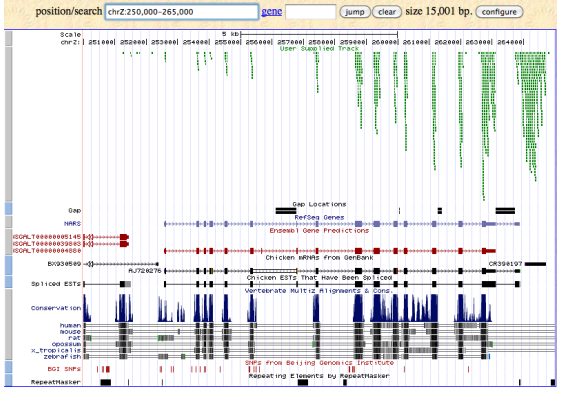
- Accurate and provably-suboptimal history reconstruction in a number of scenarios
- Relaxation methods are fast alternatives over larger networks
- Partial diffusion data is not a bottleneck
  - Missing diffusion data as well as its temporal features can be accurately estimated
- We outperform the competing approaches on both social, and environmental networks

3C Deconvolution Problem

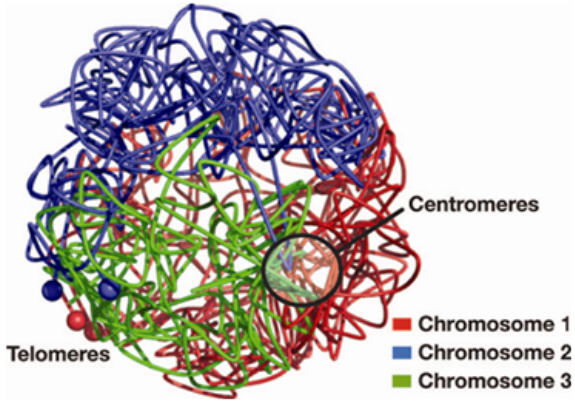
Now, let's move to a different problem.

The Spatial Organization of Genome

- 3D genome organization is linked to the cell's functional role
- Linear view restricts our understanding of the complex dynamics:
  - Long-range transcriptional control and regulation
  - eQTLs are statistically significantly closer in 3D
  - Genome folding dynamics

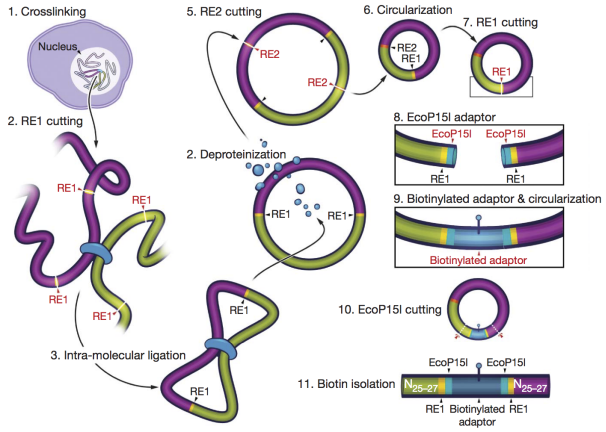


a) Linear view

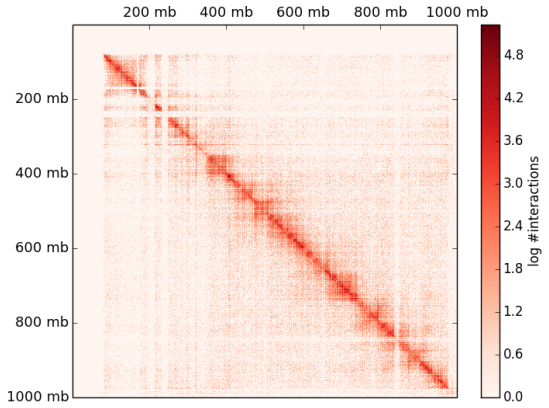


b) 3D view

Chromosome Conformation Capture (3C)



a) 3C Steps [9]



b) Interaction Matrix [10]

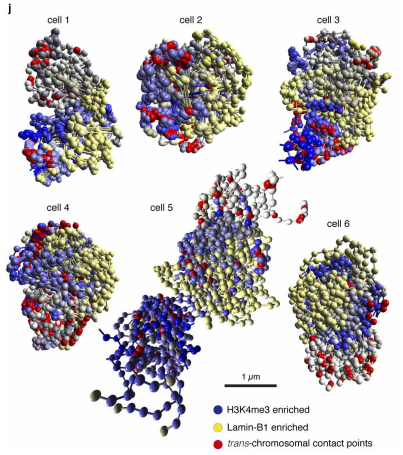
- 3C is based on restriction fragments cutting and mapping
- Raw data is binned at a given resolution
  - 0-100 kb, 100-200 kb, ...

---

[10] Duan et al., "A Three-dimensional Model of the Yeast Genome".

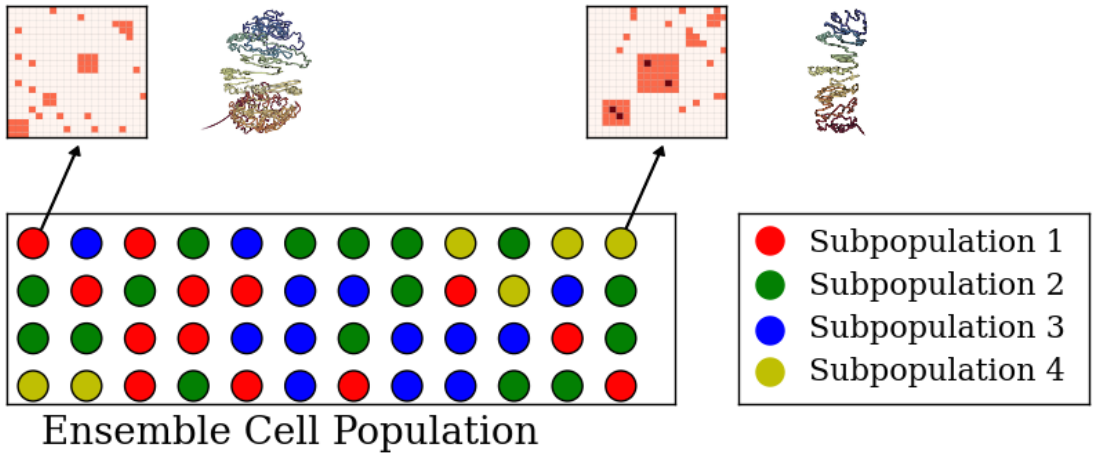
[10] Dixon et al., "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions".

The Importance of 3C-based Methods



- 3*C* data is obtained at a higher resolution
- Analyzing interaction matrix reveals interesting findings
  - 3D Embedding, Normalization
- HOXA conformation may not be fully understood without 3C

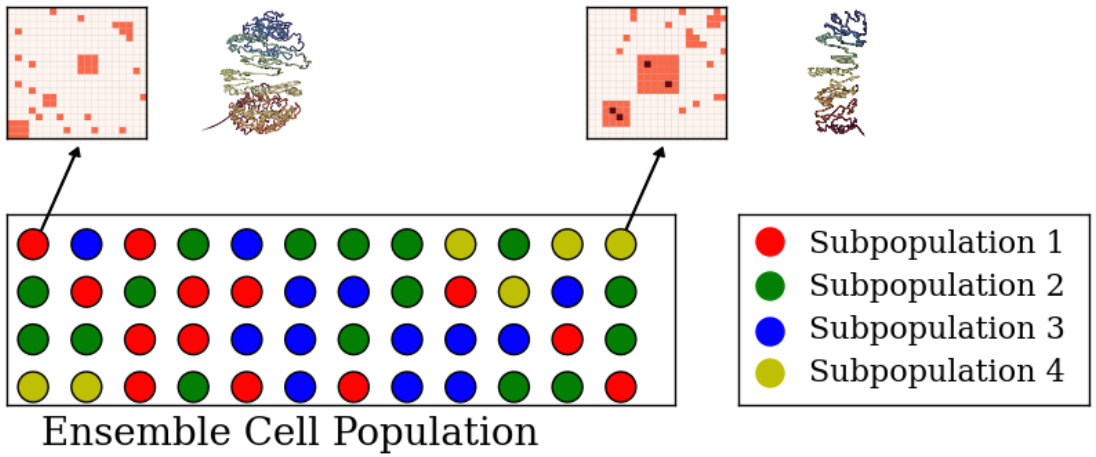
3C Deconvolution Problem



BUT

- 3C data is obtained over a cell population
  - Each cell has different interaction matrix.
- Ensemble solution is not sufficient:
  - Cells perform different functions in each phase (Temporal)
  - Each cell shows different response to the stressors (Spatial)

3C Deconvolution Problem

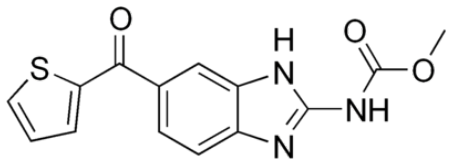


- **Main Question:** Can we unmix 3C matrices and identify meaningful latent mixing structures[11]?

[11] Emre Sefer, Geet Duggal, and Carl Kingsford. "Deconvolution Of Ensemble Chromatin Interaction Data Reveals The Latent Mixing Structures In Cell Subpopulations". In: RECOMB'15.

Why Computational Problem is Important?

- Measure the interaction matrix of each single cell[12]
  - Need multiple experiments
- Measure the interactions at a particular cell phase[13]
  - Chemicals may disrupt the genome shape

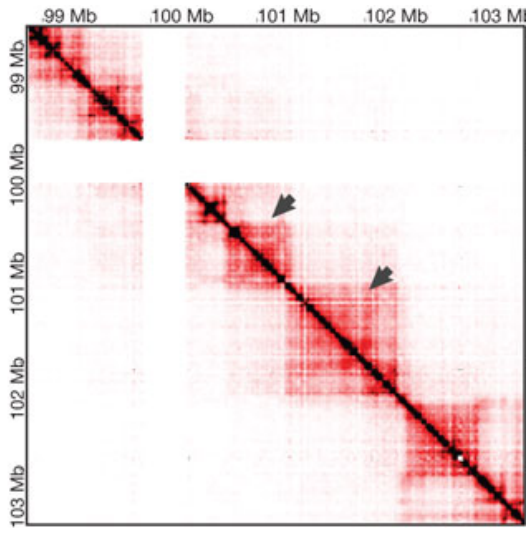


Nocodazole

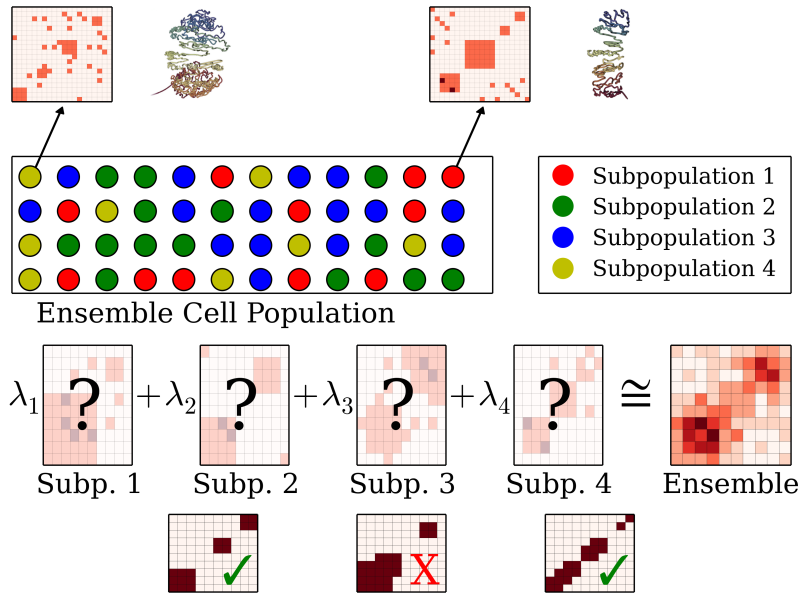
[12] Takashi Nagano et al. "Single-cell Hi-C Reveals Cell-to-cell Variability in Chromosome Structure". In: *Nature* (2013).

[13] Natalia Naumova et al. "Organization of the Mitotic Chromosome". In: *Science* (2013).

Assumptions about Deconvolution



14



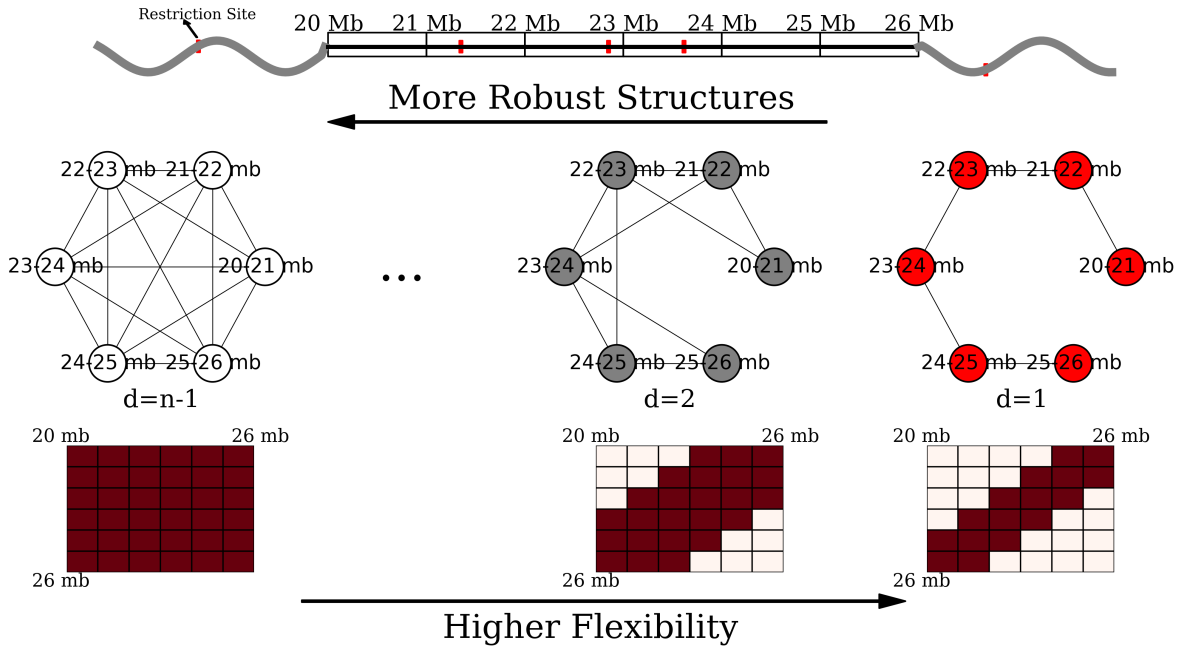
- Genome is made up of topological domains
  - highly self-interacting
  - robust consecutive genomic regions
  - building blocks
  - few megabases in length

---

[14] Nora et al., "Spatial partitioning of the regulatory landscape of the X-inactivation centre".

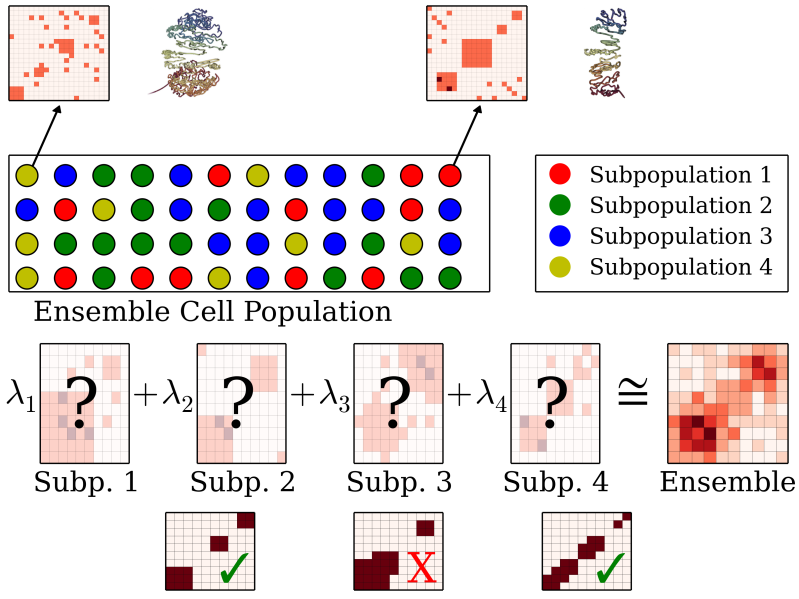
Assumptions about Deconvolution

- Flexible *Bandwidth-quasi-cliques* (*BQC*'s)



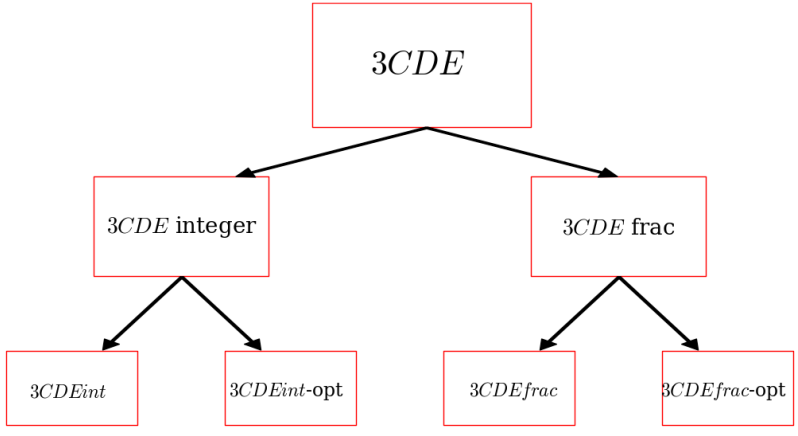
- Overlapping regions tend to disappear at a single-cell level
- Domains are independent of each other
- Domain may not need to be a clique

Frequency Deconvolution Problem *3CDE*



$$\mathbf{F} \approx \sum_i \lambda_i \mathbf{F^i}$$

Densities ↑ λ_i Matrices ↑ $\mathbf{F^i}$



Approximate Methods

$$\mathbf{F} \approx \sum_{i} \underset{\text{Densities}}{\lambda_i} \underset{\text{Matrices}}{\mathbf{F^i}}$$

Algorithm 1 Iterative two-step method for bisupermodular minimization (6)–(8)

1: $Y = \{(i, 1) | \, i \in I\}$
2: **while** there is improvement in the objective (6) **do**
3: $\quad X = \operatorname{argmin}_{A \in X,} \; Q(A, Y)$
4: $\quad Y = \operatorname{argmin}_{B \in Y,} \; Q(X, B)$
5: **end while**

Step1: Non-monotone Supermodular Independent Set in Interval Graph

$$\min Q(X|Y) = \sum_{(u,v)\in V^2} \left(F_{u,v} - \left(\sum_{i,s\in Y} s \left(\sum_{p\in M(u)\cap M(v)} \sum_{d=|u-v|}^{l_p-1} x_{pdi} \right) \right) \right)^2 + \sum_{i\in I} \sum_{(p,d)\in V_q} w^c_{pd}(1 - x_{pdi}) \quad (1)$$

$$\text{s.t} \quad x_{pdi} + x_{rti} \leq 1, \quad \forall ((p,d),(r,t)) \in E_q, \ \forall i \in I \quad (2)$$

$$x_{pdi} \in \{0,1\}, \quad \forall (p,d) \in V_q, \ \forall i \in I \quad (3)$$

1: Solve quadratic LP relaxation
2: Run specialized randomized rounding and remove the intersecting BQC's

Lemma

Step 1 can be approximated to a factor $\frac{1}{e} + (1 - \frac{1}{e})\overline{Q}$.

Step2: SDP Relaxation of Binary Least Squares for Density Assignment

- We define a variable $\forall s \in S^{'} = \{2^d \mid d \in 1, \ldots, \lfloor \log(F_{max}) \rfloor\}$
- We turn it into a boolean program

$$\min_{Y} \quad \mathbf{y''^{T}Ay''} - 2\mathbf{b}^{\mathsf{T}} r\mathbf{y''} + \|\mathbf{b}\|^2 \tag{4}$$

$$\text{s.t.} \quad y''^{2}_{is} = 1, \quad i \in 1, \ldots, k\,, s \in S' \tag{5}$$

$$r^2 = 1 \tag{6}$$

1: Solve SDP relaxation
2: Quantize each into the binary vector by their sign
3: Return $\hat{y} = \min_{l \in 1,\ldots,L} \hat{y}_l^T \mathbf{A} \hat{y}_l$.

Lemma

Step 2 can be approximated to a factor $\frac{2}{\pi} + (1 - \frac{2}{\pi})\overline{Q}$.

3CDEfrac

- We modify only the second step for fractional class densities.
- Optimally solve the convex quadratic program.

$$\min_{Y} \sum_{i \in I} \sum_{j \in I} \Big(\sum_{(u,v) \in V^2} m_{ui} m_{vj} \Big) y_i y_j \; - 2 \sum_{i \in I} \Big(\sum_{(u,v) \in V^2} F_{uv} m_{ui} m_{vi} \Big) y_i \tag{7}$$

$$y_i \geq 0, \quad i \in I \tag{8}$$

Exact Deconvolution Methods (*3CDEint*-opt, *3CDEfrac*-opt)

- *3CDEint*-opt is a convex Quadratic Integer Program (QIP).
- *3CDEfrac*-opt is Mixed Integer Quadratic Program (MIQP).
- Impossible to run the exact methods on larger datasets.

Performance Evaluation

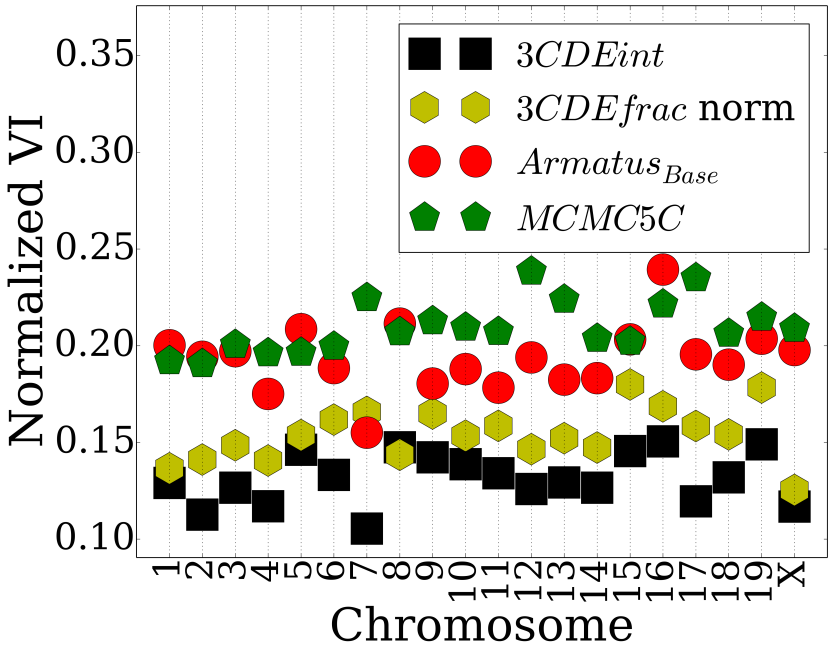
- NVI: Normalized Variation of Information
  - Measures the distance between true and estimated partitions
- MAE: Normalized Mean Absolute Error
  - Measures the absolute matrix error

$$MAE(\mathbf{T^i}, \mathbf{E^j}) = \frac{\sum_{u \in V} \sum_{v \in V} |T^i_{u,v} - E^j_{u,v}|}{n^2}$$

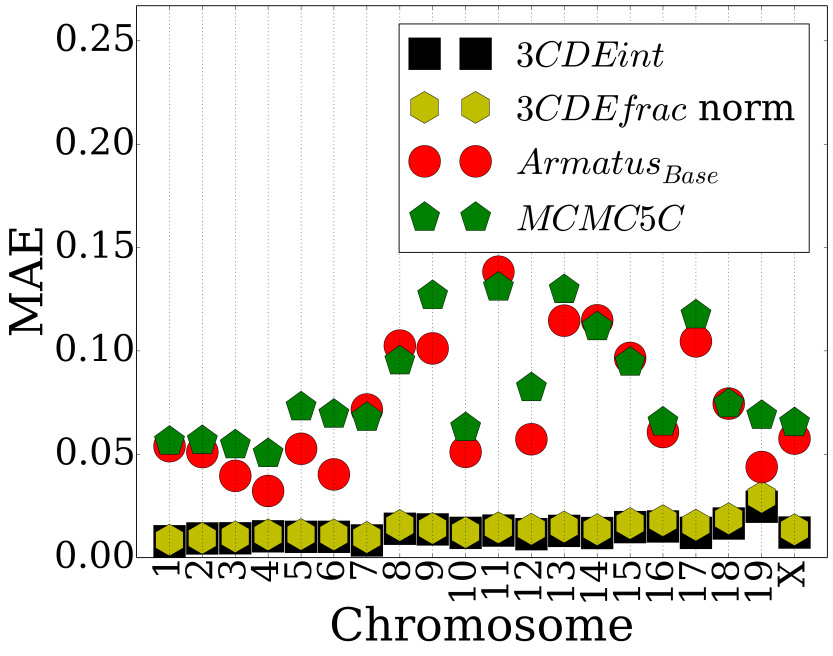
- $\text{Armatus}_{\text{Base}}$: Greedily assigns *Armatus* domains to each class without overlap

Deconvolution of Single Interaction Matrices in $CD4^+$

- We perform significantly better than the competing methods



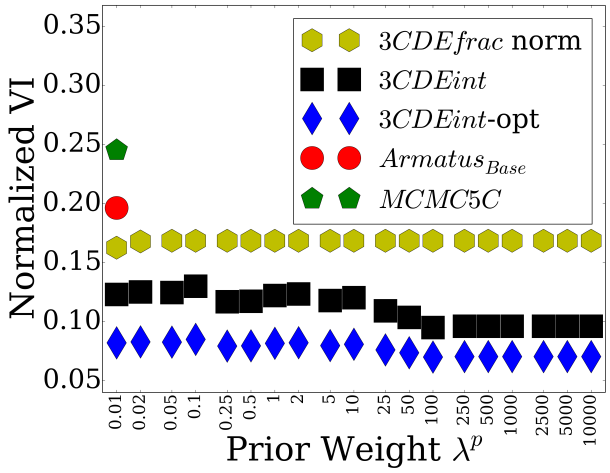
(a) Chromosome-wise NVI



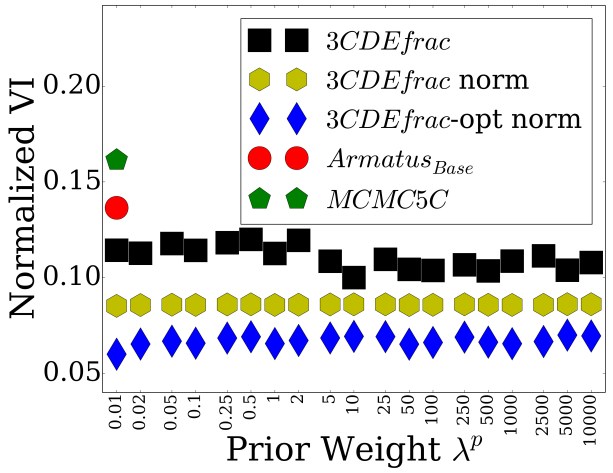
(b) Chromosome-wise MAE

Spatial and Temporal Deconvolution

- Semi-synthetic experiments in *HeLa* cellls
- Normalized data is more robust to prior weights.
- Exact methods slightly improve the performance:
  - Deconvolution problem is inherently difficult.



(a) Chro. 17

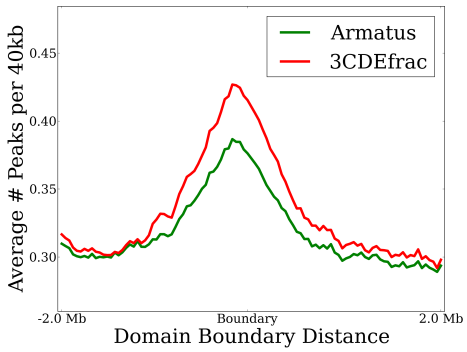


(b) HeLa

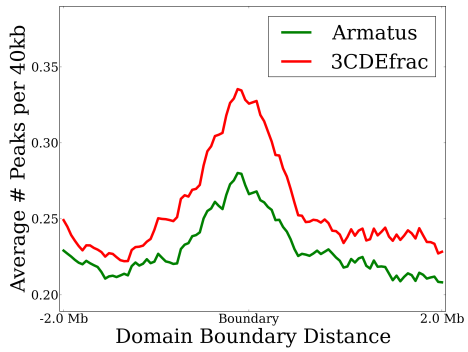
- We may also use our methods as topological domain finders.

Distribution of the Epigenetic Marks

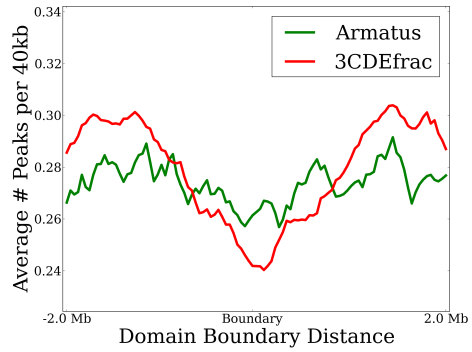
- Epigenetic marks are important in domain formation



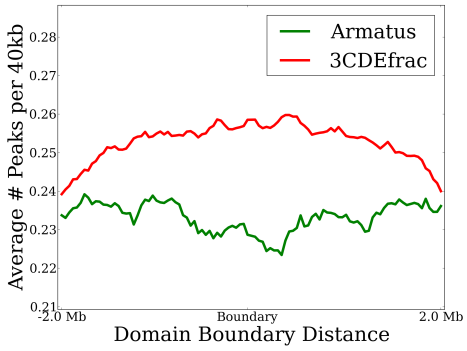
(a) H4K4me3 CD4$^+$



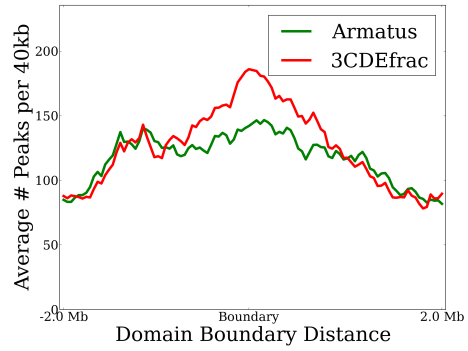
(b) H3K27ac CD4$^+$



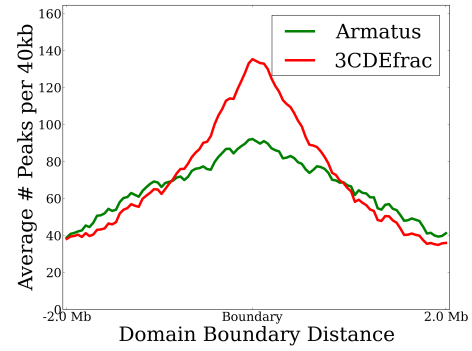
(c) H3K9me3 CD4$^+$



(d) H3K4me1 CD4$^+$



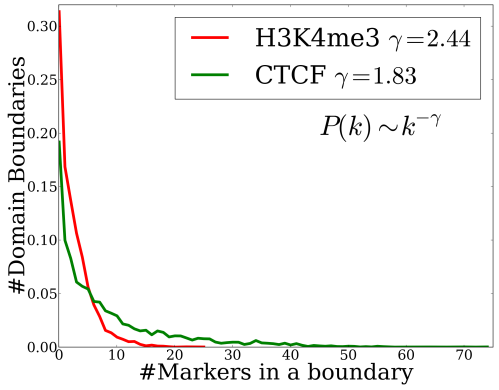
(e) H3K4me3 HeLa



(f) CTCF HeLa

Distribution of the Epigenetic Marks

- H3K4me3, H3K27ac, Pol2 are significantly enriched in the domain boundaries ($p < 0.05$, Shuffle Test)
  - Not true for distribution inside domains
  - Boundary formation is critical in domain formation
- Power-law distribution for H3K4me3 in the domain boundaries
  - Power-law exponent depends on the marker type

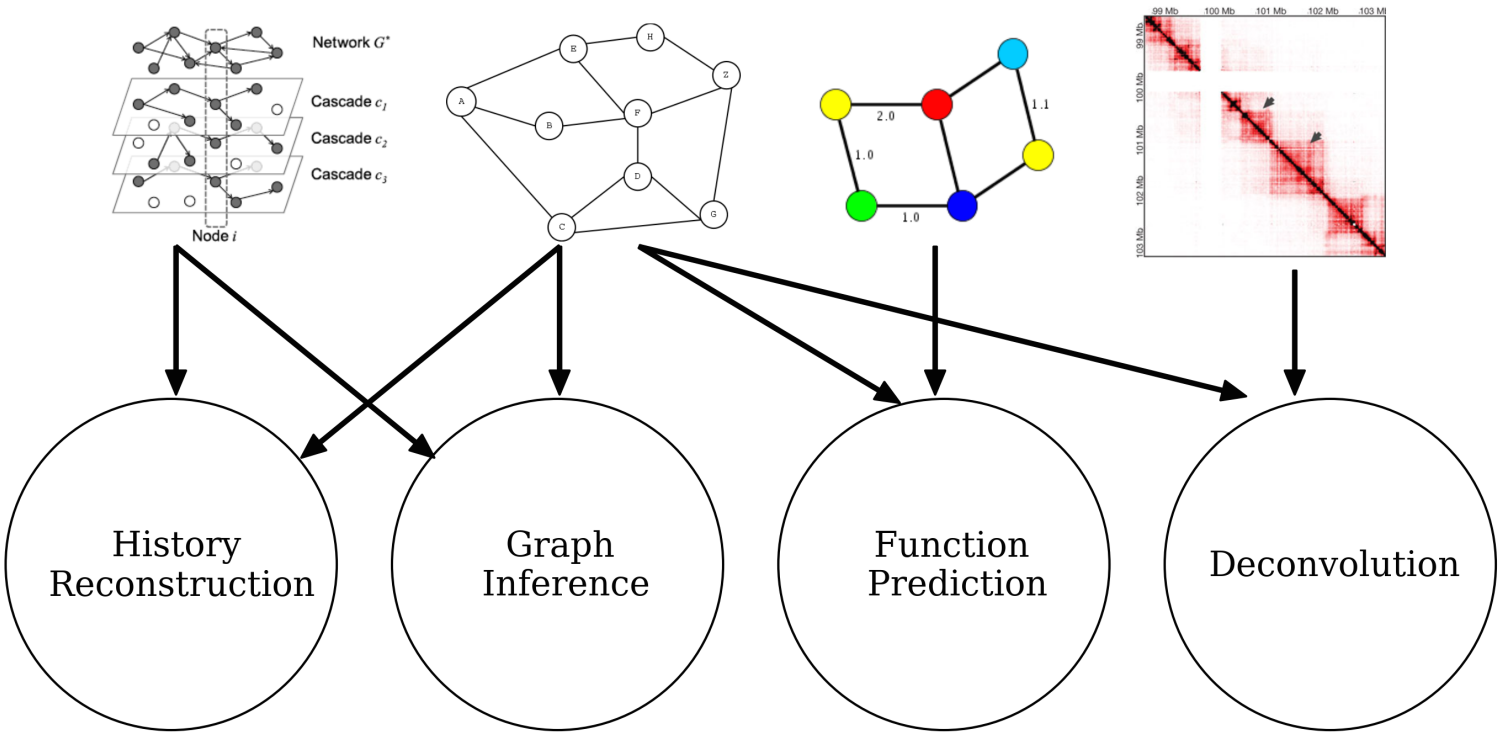


Conclusion

- We can extract the latent interaction data in 3C ensemble by deconvolution
  - Promising results even without biological priors.
- We return biologically-plausible domain decompositions
- Domain formation is related to the epigenetic marker distribution
  - Direction is still unknown.

Recap: Thesis Summary

- Improved modeling reveals the latent information in different dimensions



Future Work

- Inference and history reconstruction for arbitrary diffusion dynamics
  - Interaction between multiple cascades
  - Models for specific problems such as metastatis progression
- How do genetic and epigenetic marks affect the domain formation? (In Progress)
  - Different types of chromatin domains

Acknowledgements

- Thanks to my advisor Carl Kingsford
- Thanks to the funding

- Kingsford Group Members
  - Geet Duggal
  - Darya Flippova
  - Rob Patro
  - Hao Wang
  - Bradley Solomon
- Committee Members