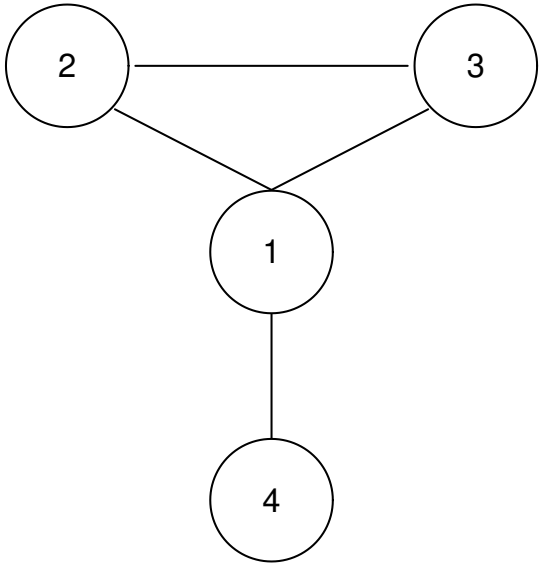
# Network Data




$$\mathbf{X} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

# Stochastic Block Models (SBMs)

- notion of "community", "group", "block", ...
- let $z_i, z_j \in \{1, 2, ..., K\}$ = group memberships of $i$ and $j$
- given $\mathbf{Z} = \{z_i : 1 \leq i \leq n\}$, SBM assumes

$$\mathcal{M}(\mathbf{X}|\mathbf{Z}) = \prod_{i,j} \left(p_{z_i z_j}\right)^{x_{ij}} \left(1 - p_{z_i z_j}\right)^{1-x_{ij}}$$

| | |
|---|---|
| observable: | $\{x_{ij} : 1 \leq i, j \leq n\}$ |
| unobservable: | $\{z_i : 1 \leq i \leq n\}$<br>$\{p_{k\ell} : 1 \leq k, \ell \leq K\}$ |

# Network Data Over Time

- discrete time:

$$\mathbf{X}^{(0)} \longrightarrow \mathbf{X}^{(1)} \longrightarrow ... \longrightarrow \mathbf{X}^{(t)} \longrightarrow ...$$

- continuous time:

| from $(i)$ | to $(j)$ | time of transaction $(t)$ |
|---|---|---|
| 1 (Amy) | 2 (Bob) | November 7, 2016, 23:42 |
| 2 (Bob) | 1 (Amy) | November 8, 2016, 07:11 |
| 2 (Bob) | 4 (Dan) | November 8, 2016, 07:37 |
| $\vdots$ | $\vdots$ | $\vdots$ |

e.g., email communication

# Non-homogeneous Poisson Process

- observe $m$ events at $t_0 < t_1 < t_2 < ... < t_m < t_\infty$
- the joint density of the arrival times is given by

$$f(t_1, t_2, ..., t_m) = \underbrace{\left( e^{-\int_{t_0}^{t_\infty} \rho(u)du} \right)}_{J(\rho)} \times \prod_{h=1}^{m} \rho(t_h)$$

  where $\rho(t)$ is the rate function of the underlying process

# SBMs for Transactional Networks

- model events (transactions) between each $(i, j)$ with a non-homogeneous Poisson process
- otherwise, inherit key features of (regular) SBMs, i.e.,
  - rate function governed by group membership, $\rho_{z_i z_j}(t)$
  - transactions conditionally independent

# The Probability Model

- given $z_i, z_j \in \{1, 2, ..., K\}$ = group memberships of $i$ and $j$

$$\mathcal{M}(\mathbf{T}|\mathbf{Z}) = \prod_{i,j} \left\{ J(\rho_{z_i z_j}) \times \prod_{h=1}^{m_{ij}} \rho_{z_i z_j}(t_{ijh}) \right\}$$

where $\mathbf{T} = \{t_{ijh} : 1 \le i, j \le n; h = 1, 2, ..., m_{ij}\}$,

$\mathbf{Z} = \{z_i : 1 \le i \le n\}$

| | |
|---|---|
| observable: | $\{t_{ijh} : 1 \le i, j \le n; h = 1, 2, ..., m_{ij}\}$ |
| unobservable: | $\{z_i : 1 \le i \le n\}$<br>$\{\rho_{k\ell} : 1 \le k, \ell \le K\}$ |

# Basketball Games

| From | To | Time | "Active" Teammates |
|---|---|---|---|
| inbound | C#9 | 0 | C#9, C#20, C#30, C#34 |
| C#9 | C#5 | 11 | C#5, C#9, C#20, C#30, C#34 |
| C#5 | miss 2 | 12 | C#5, C#9, C#20, C#30, C#34 |
| rebound | H#6 | 0 | H#3, H#6, H#15, H#21, H#31 |
| H#6 | H#3 | 7 | H#3, H#6, H#15, H#21, H#31 |
| H#3 | H#15 | 8 | H#3, H#6, H#15, H#21, H#31 |
| H#15 | H#3 | 9 | H#3, H#6, H#15, H#21, H#31 |
| H#3 | H#6 | 12 | H#3, H#6, H#15, H#21, H#31 |
| H#6 | miss 3 | 17 | H#3, H#6, H#15, H#21, H#31 |

C = Boston Celtics; H = Miami Heat

# Model Adaptation/Extension

- considered three initial states,

$$\mathcal{S} = \{\text{inbound, rebound, steal}\}$$

and six absorbing states,

$$\mathcal{A} = \{\text{make 2, miss 2, make 3, miss 3, fouled, turnover}\}$$

- partitioned the probability model into three parts:

$$\underbrace{\left[\prod_{s\in\mathcal{S}}\prod_{i=1}^{n}\cdots\right]}_{\text{initial}} \times \underbrace{\left[\prod_{i,j}\cdots\cdots\right]}_{\text{passing}} \times \underbrace{\left[\prod_{j=1}^{n}\prod_{a\in\mathcal{A}}\cdots\right]}_{\text{outcome}}$$

with transition probabilities $p_{sk}$, $p_{ka}$ for all $a \in \mathcal{A}$, $s \in \mathcal{S}$

# ... will "gloss over" many tricky details ...

Reference

Xin L, Zhu M, Chipman HA (2017), "A continuous-time stochastic block model for basketball networks," *The Annals of Applied Statistics*, vol. 11, no. 2, pp. 553–597.

# Some Key Modifications

$$\mathcal{M}(\mathbf{T}|\mathbf{Z}) = \prod_{i,j} \left\{ J(\rho_{z_i z_j}) \times \prod_{h=1}^{m_{ij}} \rho_{z_i z_j}(t_{ijh}) \right\}$$

- modification (a):

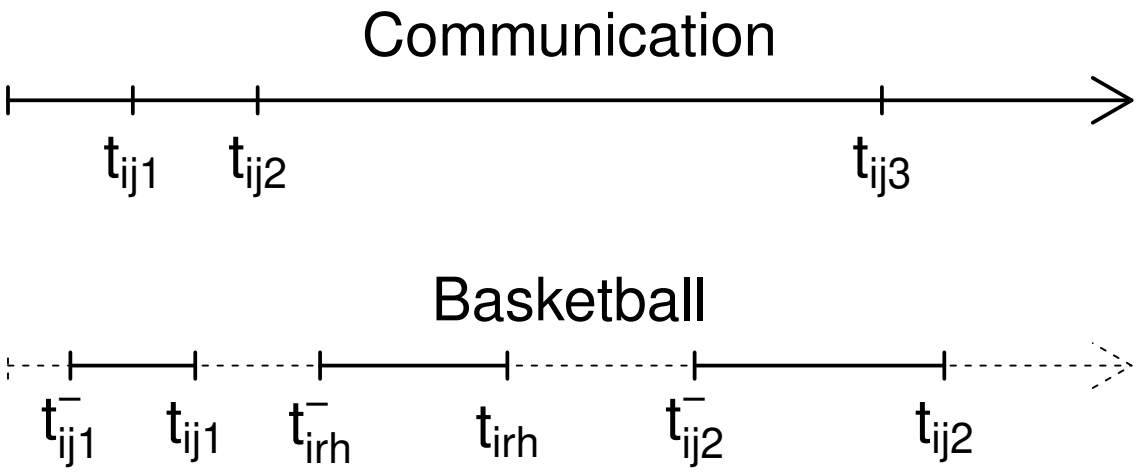
$$\rho_{z_i z_j}(t_{ijh}) \quad \Longrightarrow \quad \rho_{z_i z_j}(t_{ijh}) \times \frac{1}{G_{z_j}^{ijh}},$$

  where $G_{z_j}^{ijh}$ = # of "eligible receivers" in group $z_j$ for the $h$-th pass between $i$ and $j$, because $\exists$ only ONE ball

- modification (b):

$$J(\rho_{z_i z_j}) \quad \Longrightarrow \quad J_{ij}(\rho_{z_i z_j}),$$

  also because $\exists$ only ONE ball $\longrightarrow$ see next slide




$$J(\rho_{z_i z_j}) \equiv \exp\left[-\int_{t_0}^{t_\infty} \rho_{z_i z_j}(t)dt\right] \quad \Longrightarrow$$

$$J_{ij}(\rho_{z_i z_j}) \equiv \exp\left[-\sum_{r\neq i}\sum_{h=1}^{m_{ir}} \int_{t^-_{irh}}^{t_{irh}} \rho_{z_i z_j}(t) \times \frac{\boldsymbol{I}_j^{irh}}{G_{z_j}^{irh}} dt\right]$$

where $\boldsymbol{I}_j^{irh} = 1$ if player $j$ is an "eligible receiver" for the $h$-th pass between $i$ and $r$, and $= 0$ otherwise

# Further Simplification

- re-parameterized rate function $\rho_{k\ell}(t)$ as

$$\rho_{k\ell}(t) = \lambda_k(t) \times p_{k\ell},$$

where

$$\begin{aligned} \lambda_k(t) &= \text{rate function for the ball ``leaving'' group } k, \\ p_{k\ell} &= \mathbb{P}(\text{group } k \text{ passes to group } \ell) \end{aligned}$$

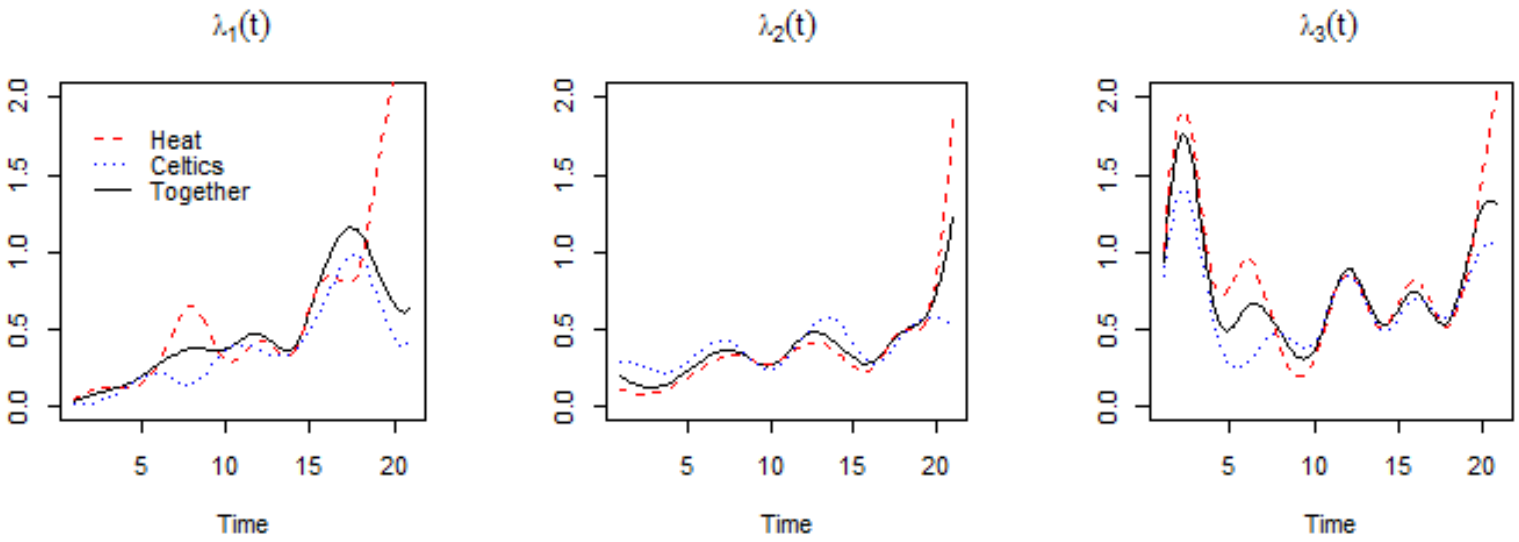
- scalar parameters $\{p_{sk}, p_{k\ell}, p_{ka}\}$ all have quick closed-form updates in our iterative algorithm

- whereas functional parameters $\{\lambda_k(t)\}$ require quasi-Newton updates

# Case Studies

- two games (Game 1 and Game 5) from the 2012 Eastern Conference finals between the Miami Heat and the Boston Celtics
- two games (Game 2 and Game 5) from the 2015 NBA finals between between the Cleveland Cavaliers and the Golden State Warriors

# 2012 Heat vs Celtics




$$g_1 = \{\text{point guards}\}; \qquad g_3 = \{\text{others}\}$$

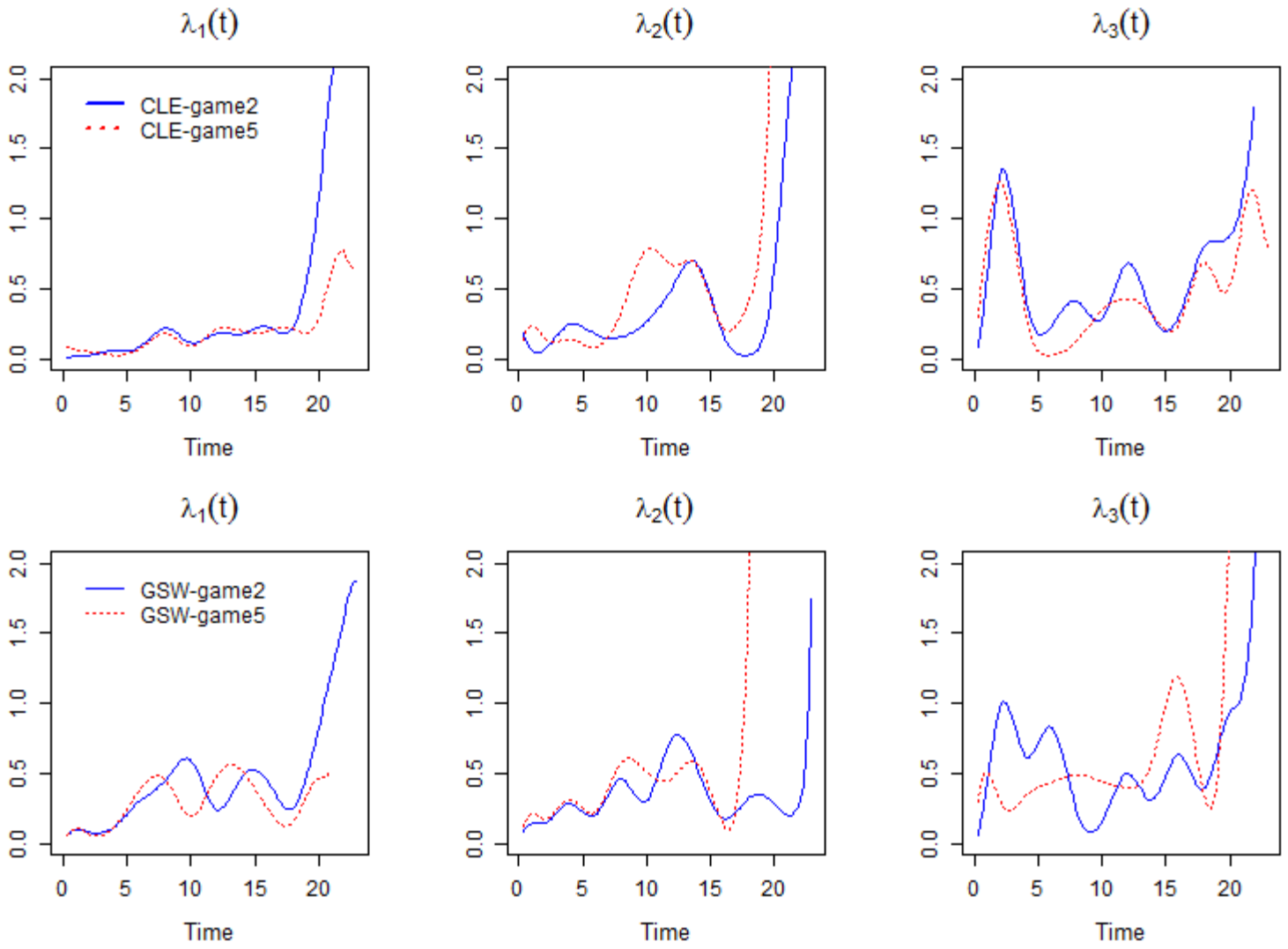
$$g_2 = \{\text{Heat: Dwyane Wade + LeBron James}\}$$
$$\{\text{Celtics: Ray Allen + Paul Pierce}\}$$

Remarks

(i) Celtics: Brandon Bass $\in g_3$, but $\in g_2$ if analyzed w/ Heat players.

(ii) Celtics: Rajon Rondo is the "cause" of "unusual" $\lambda_1(t)$, $\lambda_3(t)$.

# 2015 Cavaliers vs Warriors



# 2015 Cavaliers vs Warriors

| Game | $g_1$ | $g_2$ | $g_3$ |
|---|---|---|---|
| #2 | PGs<br>Andre Iguodala (SF)<br>Draymond Green (PF) | SGs | SFs<br>Centers |
| #5 | PGs | SGs<br>SF+PF | Andre Iguodala (SF)<br>Draymond Green (PF) |

PG = point guard (Stephen Curry + Shaun Livingston)
SG = shooting guard (Klay Thompson + Leandro Barbosa)
SF = shooting forward (Harrison Barnes)
PF = power forward (David Lee)

Remark

In the 2015 final, the Golden State Warriors famously changed their lineup after losing games 2 & 3 and went on to win the championship.

# LeBron James

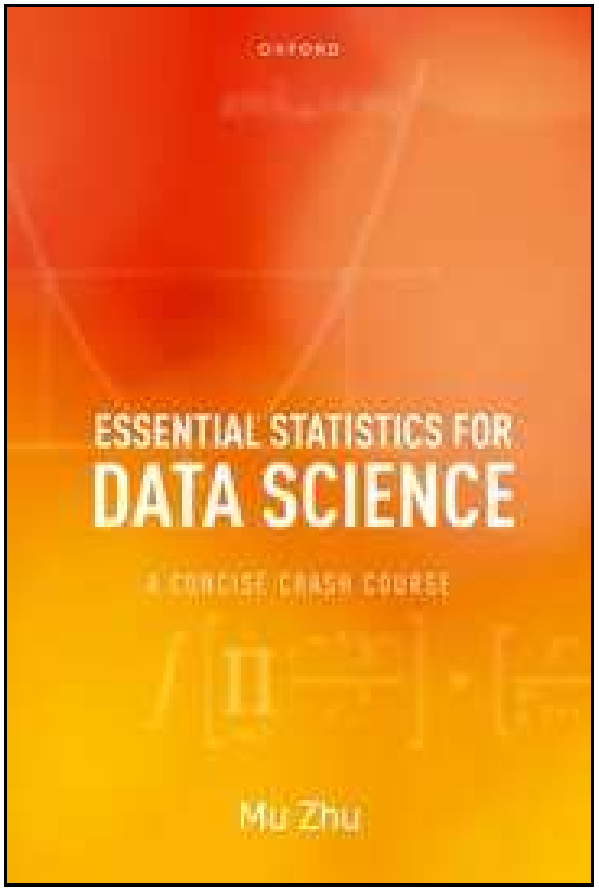
- on 2011/12 Miami Heat + 2014/15 Cleveland Cavaliers
- grouping results (more players being analyzed, $K = 4$):
  - $g_1$: {point guards}
  - $g_2$: {LeBron James$^{(H)}$, Dwyane Wade, LeBron James$^{(C)}$}
  - $g_3$: {other perimeter players}
  - $g_4$: {big men, i.e., power forwards + centers}
- in fact, experts have long suggested the need to create another on-court position — e.g., "Point Forward" — due to his distinctive playing style

# Statistical Approach to Data Science

- set up probabilistic model
  - to describe data generating process
- learn the probabilistic model
  - by estimating its parameters (and other unknowns)
- the estimated model can then be used
  - to reveal patterns
  - to gain insights
  - to make predictions

**Remark** In the (vanilla or continuous-time) SBMs, the "other unknowns" — namely, $z_i, z_j$ — are called "latent variables".

# Reference




Zhu M (2023), *Essential Statistics for Data Science: A Concise Crash Course*, Oxford University Press.

Chapter 1
Eminence of Models

https://doi.org/10.1093/oso/9780192867735.003.0001