O Estudo de Processos Pontuais por meio da Análise Topológica de Dados e Aplicações

aluno: Rafael Polli Carneiro
orientador: Cristian Coletti

MEMBROS DA BANCA:
Presidente: Cristian Favio Coletti - UFABC
Membro Titular: Daniel Miranda Machado - UFABC
Membro Titular: Natalia Andrea Viana Bedoya - UFSCAR
Membro Suplente: Erika Alejandra Rada Mora - UFABC
Membro Suplente: Mariana Rodrigues Da silveira - UFABC

1. Introdução

Introdução

Estes slides consistem em apresentar meus resultados preliminares para os seguintes problemas

1. Primeiro Problema:

   Estudar "propriedades persistentes" de uma coleção de pontos oriundos de um Processo Pontual não estacionário;

2. Segundo Problema:
   - ***Avaliar o comportamento de um grafo de conversas referentes a COVID-19 no Twitter***
   - ***e estudar as "propriedades persistentes" de seu subgrafo formado pelas mensagens classificadas como Fake News;***

Introdução

- O termo "**propriedades persistentes**" é destacado de propósito;
- De fato, este termo vem por causa da **Homologia Persistente**;
- E será objeto central de toda minha argumentação;
- Para isso, antes de prosseguirmos
- Precisamos entendê-la
- E criar intuição sobre a ferramenta central para nosso trabalho...

O que é Homologia Persistente

- Como dito no slide anterior
- Por **propriedades persistente** nos referimos a todo e qualquer elemento do ramo da

  Homologia Persistente
- Ramo este que se fundamenta fortemente na **Homologia Simplicial**

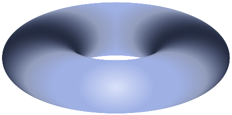
Hom. Simplicial $\Rightarrow$ Hom. Persistente

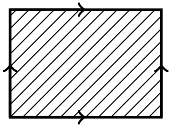
O que é Homologia Persistente

A Homologia Simplicial

- Com um olhar mais ingênuo
- Para desenvolvermos intuição
- Pensaremos na **Homologia Simplicial** como
  - uma ferramenta capaz de detectar *certas propriedades geométricas* de um conjunto
  - Tais Propriedades Geométricas serão:

componentes conexas ou “buracos”



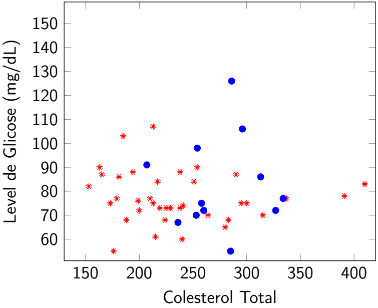


$[0, 2\pi] \times [0, 2\pi]$

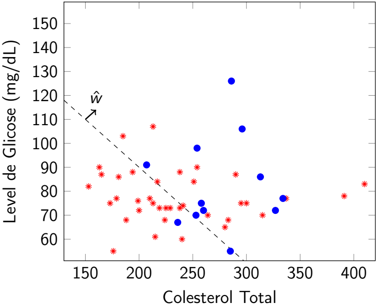
O que é a Homologia Persistente

Motivação

- Técnicas de análise estatística baseiam-se quase certamente em
  - Ajustar um hiperplano a uma amostra
  - Seguindo certos critérios de minimização



(a) Valores da amostra de 50 indivíduos.

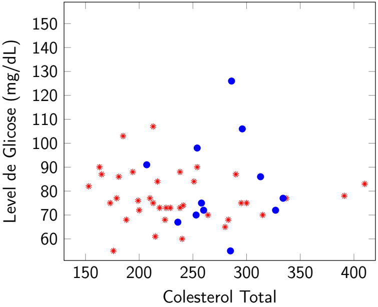


(b) Exemplo de hiperplano separando os pontos.

O que é a Homologia Persistente

Motivação

- Lado negativo destas técnicas:
  1. Propriedades locais da amostra são perdidos
  2. Em prol de se resolver um problema de minimização
- Observe-se como os pontos azuis estão mais distantes entre si;
- Se comparados aos pontos vermelhos;



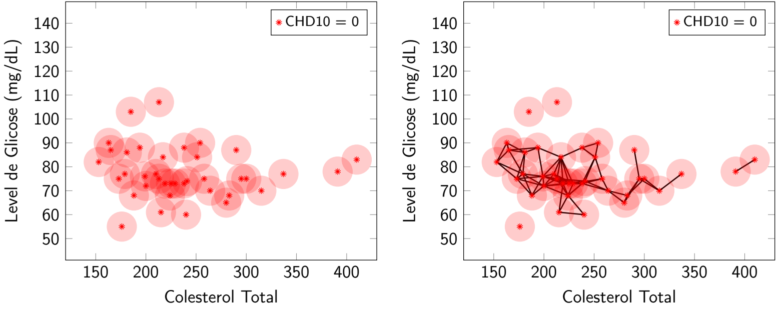
O que é a Homologia Persistente

Motivação

- **Observações**
  - Seria legal incorporar propriedades locais no modelo de análise
  - E também seria legal avaliar as características geométricas da amostra
- Motivados por estes pontos, as seguintes ideais parecem razoáveis
  1. Agrupar os pontos distantes entre si, $d(x, y) \leq r$
  2. E aplicar a homologia simplicial nestes agrupamentos (simplexos)

O que é a Homologia Persistente

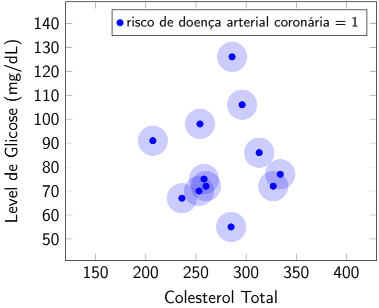
Motivação



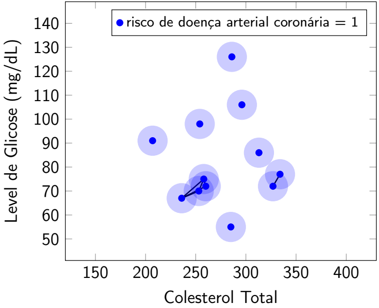
(c) (d)

O que é a Homologia Persistente

Motivação



(e)



(f)

O que é a Homologia Persistente
Motivação

- Até agora tudo bem...
- Porém:

  Qual valor r usar para agrupar os pontos em função de suas distâncias
  $$d(x,y) \leq r???$$
- Aí que surge a

Homologia Persistente

O que é a Homologia Persistente

Finalmente, a Homologia Persistente

- Seja X um conjunto finito (a amostra);
- E X_r uma sequência indexada por $r \in \mathbb{R}_{\geq 0}$:

$$X_r = \{\sigma \subseteq X; x, y \in d(x, y) \leq r\}$$

- Onde

$$X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots$$

- Temos que

$$\{X_r\}_{r \in \mathbb{R}_{\geq 0}} \text{ será chamada de filtração}$$

- Da propriedade $X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots$
- A filtração nos garante o seguinte diagrama comutativo

O que é a Homologia Persistente

Finalmente, a Homologia Persistente

$$
\begin{tikzcd}
 & \vdots & \vdots & \vdots & \\
\cdots \arrow[r] & H_2(X_a) \arrow[r, "\iota_{*,2}^{a,b}"] \arrow[d, "\partial_2"'] & H_2(X_b) \arrow[r, "\iota_{*,2}^{b,c}"] \arrow[d, "\partial_2"'] & H_2(X_c) \arrow[r] \arrow[d, "\partial_2"'] & \cdots \\
\cdots \arrow[r] & H_1(X_a) \arrow[r, "\iota_{*,1}^{a,b}"] \arrow[d, "\partial_1"'] & H_1(X_b) \arrow[r, "\iota_{*,1}^{b,c}"] \arrow[d, "\partial_1"'] & H_1(X_c) \arrow[r] \arrow[d, "\partial_1"'] & \cdots \\
\cdots \arrow[r] & H_0(X_a) \arrow[r, "\iota_{*,0}^{a,b}"] \arrow[d, "\partial_0"'] & H_0(X_b) \arrow[r, "\iota_{*,0}^{b,c}"] \arrow[d, "\partial_0"'] & H_0(X_c) \arrow[r] \arrow[d, "\partial_0"'] & \cdots \\
 & (0) & (0) & (0) &
\end{tikzcd}
$$

$$\leq \quad a \quad \leq \quad b \quad \leq \quad c \quad \leq$$

O que é a Homologia Persistente

Finalmente, a Homologia Persistente

- Considerando cada grupo de homologia um $\mathbb{K}$-espaço vetorial

Definição

A ***Homologia Persistente*** de uma filtração $\mathcal{F} = (X_r)_{r\in\mathbb{R}_{\geq 0}}$ é nada mais que a sequência de ***módulos de persistência*** $H_p(\mathcal{F})$, com $p \in \mathbb{N} \cup \{0\}$

sendo que

Definição

Um ***módulo de persistência*** $H_p(\mathcal{F})$, é a tupla

$$H_p(\mathcal{F}) = \left(H_p(X_r), (\iota_{*,p}^{r,r'} : H_p(X_r) \to H_p(X_{r'})\right)_{r,r'\in\mathbb{R}_{\geq 0}, r\leq r'}.$$

O que é a Homologia Persistente

Finalmente, a Homologia Persistente

- Seja $H_p(\mathcal{F})$ um módulo de persistência de dimensão p
- E $\mathcal{F} = \{X_r\}_{r \in \mathbb{R}_{\geq 0}}$ uma filtração de um conjunto finito X
- Então $H_p(\mathcal{F})$ pode ser visto como o multiset

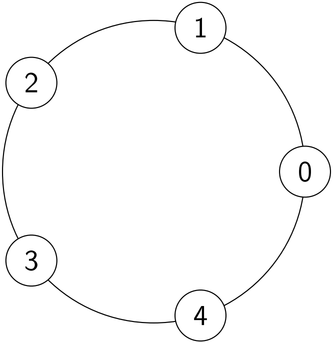
$$\{(b, d); b \leq d \in \mathbb{R}_{\geq 0}\} \subseteq \mathbb{R}^2$$

- onde $(b, d) \in \mathbb{R}^2$ representa
  - $b \rightarrow$ o instante do surgimento de um *ciclo* em X_b
  - $d \rightarrow$ o instante em que o *ciclo* se torna um *bordo* em X_d

O que é a Homologia Persistente

Um exemplo

Consideremos o grafo não orientado

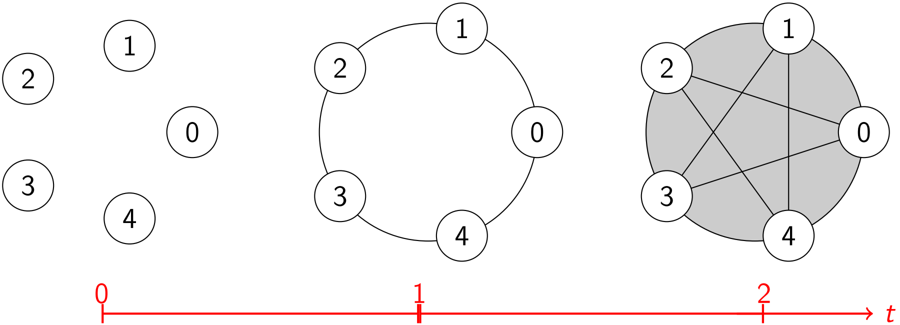


Munido da métrica do menor caminho entre vértices

O que é a Homologia Persistente

Um exemplo

O qual admite a filtração



O que é a Homologia Persistente

Um exemplo

- E os seguintes módulos de persistência

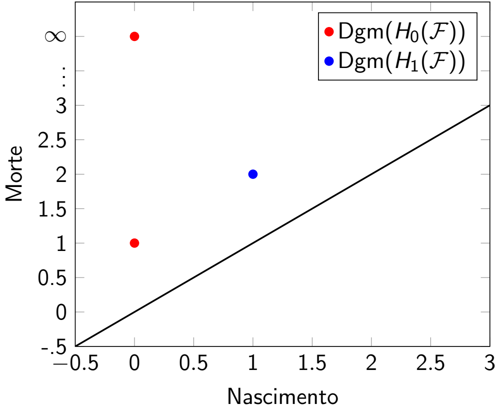
$$
\begin{array}{llllllll}
 & \mathbb{R}^5 & \xmapsto{i_{*,0}^{0,1}} & \mathbb{R} & \xmapsto{i_{*,0}^{1,2}} & \mathbb{R} & \xmapsto{i_{*,0}^{2,3}} & \mathbb{R} \quad \cdots \\
H_0(\mathcal{F}) : & 0 & & 1 & & 2 & & 3 \quad \cdots \quad \mathbb{R}_{\geq 0}
\end{array}
$$

$$
\begin{array}{llllllll}
 & (0) & \xmapsto{i_{*,1}^{0,1}} & \mathbb{R} & \xmapsto{i_{*,1}^{1,2}} & (0) & \xmapsto{i_{*,1}^{2,3}} & (0) \quad \cdots \\
H_1(\mathcal{F}) : & 0 & & 1 & & 2 & & 3 \quad \cdots \quad \mathbb{R}_{\geq 0},
\end{array}
$$

O que é a Homologia Persistente

Um exemplo

- A homologia persistente desta filtração pode ser vista pelo gráfico de nascimento e morte de ciclos



O que é a Homologia Persistente

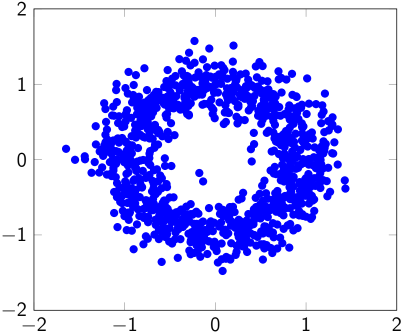
Implementação Computacional

- A vantagem de se trabalhar com filtrações simpliciais é a de existir um algoritmo simples para o cálculo da Homologia Persistente
- De fato, o algoritmo consiste basicamente em realizar uma eliminação gaussiana dos operadores bordos
- Apresentamos outro Exemplo para ilustrar que a Homologia Persistente nos fornece resultados condisentes com o esperado
- Este exemplo será resultado da eliminação gaussiana

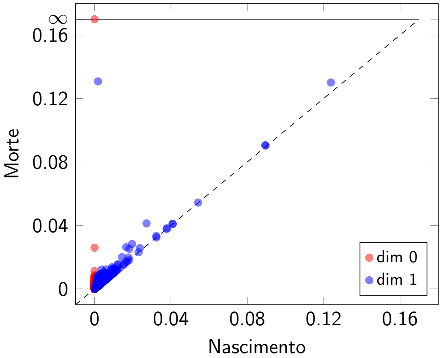
O que é a Homologia Persistente

Último Exemplo

Simulação e Cálculo da Homologia Persistente



(g)



(h)

2. Resultados - Pt1

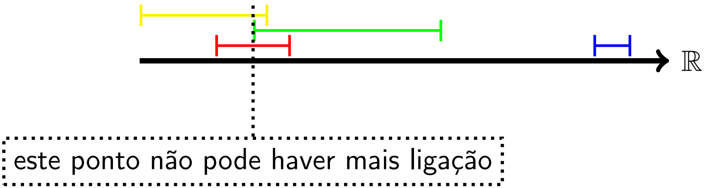
Primeira Parte do Trabalho

O Processo Pontual Estudado

- O Processo Pontual que servirá de base para nosso estudo pode ser entendido como

  A evolução temporal de chamadas telefônicas na reta real sendo que apenas C chamadas podem ocorrer simultâneamente no mesmo ponto

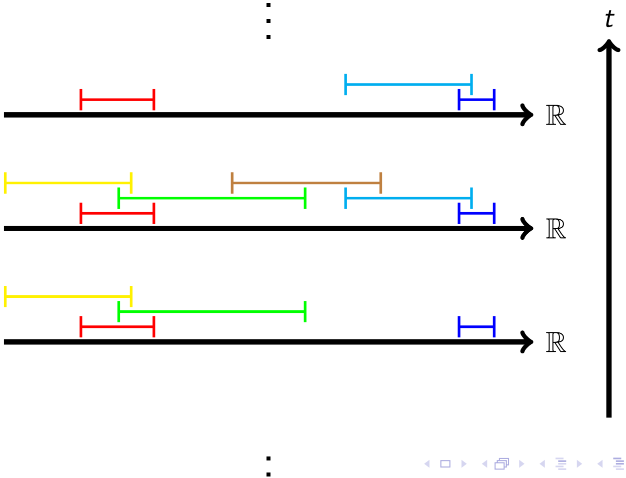
- Exemplo: fixado um tempo $t \in \mathbb{R}$ e $C = 3$ temos



Primeira Parte do Trabalho

O Processo Pontual Estudado

- O mesmo Exemplo, agora considerando sua dinâmica no tempo



Primeira Parte do Trabalho

O Processo Pontual Estudado

- Sob certas condições esta dinâmica de "nascimento" e "morte" converge para um "equilíbrio",
- ou melhor dizendo, converge

  para uma **medida invariante**

- Ou seja, temos um padrão para o Processo Pontual das chamadas telefônicas, quando $t \to \infty$
- Ai que chegamos aonde queriamos:

  Simulação Perfeita

Primeira Parte do Trabalho

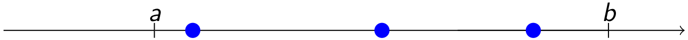
O Processo Pontual Estudado

- A **Simulação Perfeita** será a ferramenta que nos fornecerá o Processo Pontual em equilíbrio
- De forma simplificada, ela consiste em
- Dado um intervalo $[a, b]$
- E considerando que o comprimento de ligações vem de uma distribuição π

$$\pi \sim \mathcal{U}(0, H)$$

1. Gerar pontos de um Processo Pontual de Poisson de taxa λ em $[a, b]$:

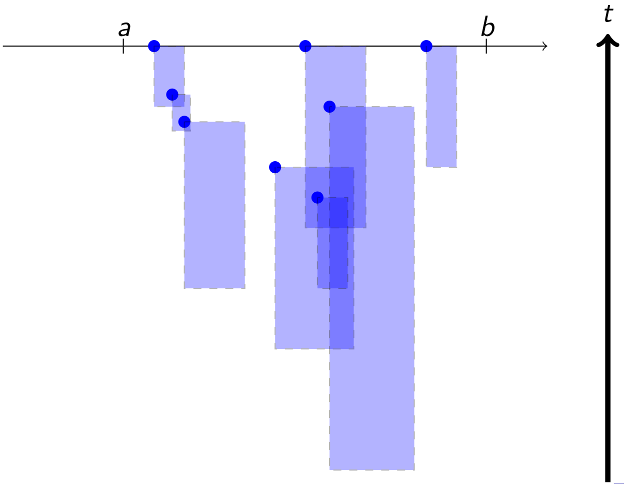
$$x_1, x_2, \ldots, x_N$$



Primeira Parte do Trabalho

O Processo Pontual Estudado

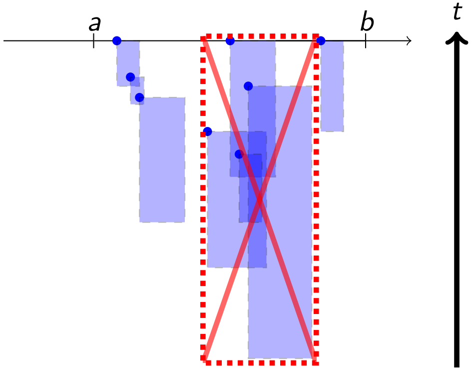
2. E gerar uma sequência de pontos ancestrais a x_i



Primeira Parte do Trabalho

O Processo Pontual Estudado

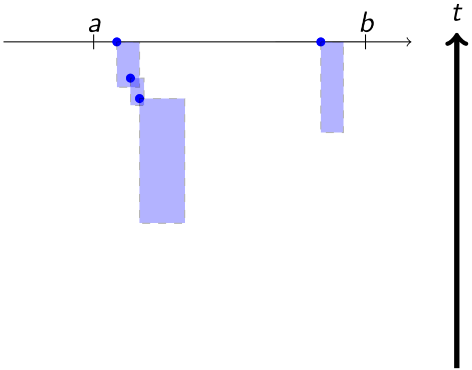
3. Com os pontos ancestrais uma filtração é realizada baseada no máximo de ligações simultâneas C
4. Por exemplo, sendo $C = 3$



Primeira Parte do Trabalho

O Processo Pontual Estudado

5. Após a filtração, o Processo Pontual será:
   - (i) Os pontos $x_1, x_2, \ldots, x_N$ restantes da filtração;
   - (ii) E os retângulos ancestrais, formando o ***clã de ancestrais***



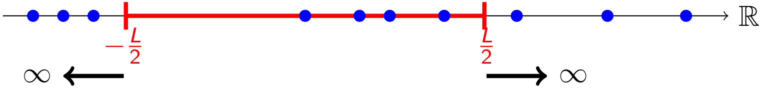
Primeira Parte do Trabalho

O Processo Pontual Estudado

- Dado o Processo Pontual resultante da Simulação Perfeita, denotado por ζ
- Vamos Estudar o comportamento da taxa do ***número de persistência de Betti de dimensão*** 0 de ζ na janela

$$\Lambda = \left[-\frac{L}{2}, \frac{L}{2}\right] \cap \zeta,$$

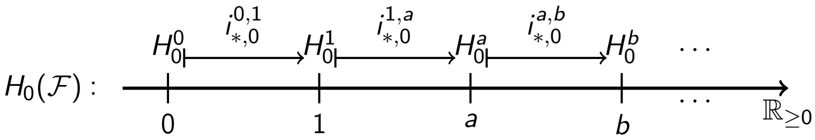
com $L \in \mathbb{R}_{>0}$, fazendo $L \to \infty$



Primeira Parte do Trabalho

O que é o Número de Persistência de Betti

- Lembremos que o módulo de Persistência de dimensão 0 será a tupla



Definição

Fixados $0 \leq a \leq b \in \mathbb{R}$, o ***Número de Persistência de Betti*** de dimensão 0, no intervalo (a, b), é o valor

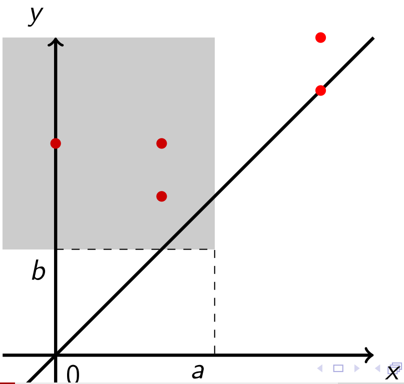
$$\beta_0^{a,b} = \dim(\mathrm{Im}(i_{*,0}^{2,3}))$$

Primeira Parte do Trabalho

O Número de Persistência de Betti como uma medida

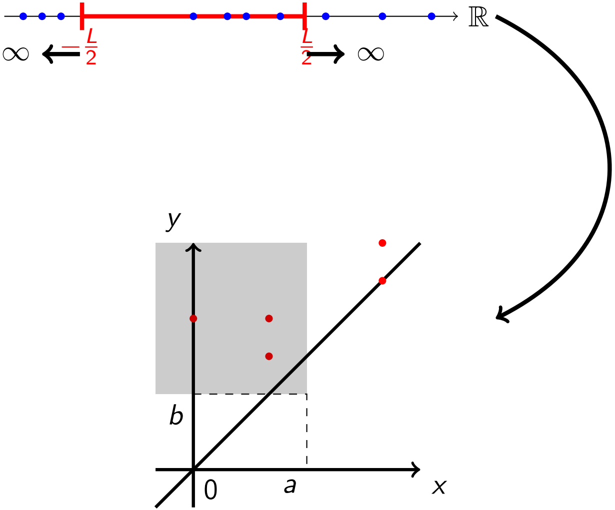
- Podemos interpretar o Número de Persistência de Betti como a quantidade de pontos do Diagrama de Persistência no retângulo

$$[-\infty, a] \times [b, \infty]$$



Primeira Parte do Trabalho

Tática



$$\frac{\beta_0^{a,b}}{L}$$

Primeira Parte do Trabalho

Resultados

Proposição

Seja ζ o processo pontual dos germes induzido pela medida invariante do processo pontual das redes com perdas, $\Lambda_L = [-L/2, L/2) \subseteq \mathbb{R}$ um intervalo e $\zeta_{\Lambda_L} = \zeta \cap \Lambda_L$. Então, considerando a filtração simplicial $\mathbb{K}(\zeta_{\Lambda_L})$ dos pontos de ζ_{Λ_L}, temos que, para todos reais $r \leq s$, existe o valor $\beta_0^{r,s}$ tal que

$$\lim_{L\to\infty} \frac{\mathbb{E}[\ \beta_0^{r,s}(\mathbb{K}(\zeta_{\Lambda_L}))\ ]}{L} = \hat{\beta}_0^{r,s}$$

Primeira Parte do Trabalho

Resultados

Proposição

Seja ζ o processo pontual dos germes induzido pela medida invariante do processo pontual das redes com perdas, $\Lambda_L = [-L/2, L/2) \subseteq \mathbb{R}$ um intervalo e $\zeta_{\Lambda_L} = \zeta \cap \Lambda_L$. Então, sempre quando o processo ζ satisfazer as condições de ergodicidade, e considerando a filtração simplicial $\mathbb{K}(\zeta_{\Lambda_L})$ dos pontos de ζ_{Λ_L}, temos que, para todos reais $r \leq s$, existe o valor $\hat{\beta}_0^{r,s}$ tal que

$$\lim_{L\to\infty} \frac{\beta_0^{r,s}(\mathbb{K}(\zeta_{\Lambda_L}))}{L} = \hat{\beta}_0^{r,s}, \quad q.c.$$

Primeira Parte do Trabalho

Resultados

Proposição

Seja ζ o processo pontual dos germes induzido pela medida invariante do processo pontual das redes com perdas, $\Lambda_L = [-L/2, L/2) \subseteq \mathbb{R}$ um intervalo e $\zeta_{\Lambda_L} = \zeta \cap \Lambda_L$. Então, considerando a filtração simplicial $\mathbb{K}(\zeta_{\Lambda_L})$ dos pontos de ζ_{Λ_L}, temos que, para todos reais $r \leq s$, existe o valor $\hat{\sigma}_{r,s}$ tal que

$$\frac{\beta_0^{r,s}(\mathbb{K}(\zeta \cap \Lambda_L)) - \mathbb{E}\left[\beta_0^{r,s}(\mathbb{K}(\zeta \cap \Lambda_L))\right]}{L^{1/2}} \xrightarrow{L \to \infty} \mathcal{N}(0, \hat{\sigma}_{r,s}^2), \qquad (1)$$

onde a convergência acima é a convergência de distribuição e $\mathcal{N}(0, \hat{\sigma}^2)$ é a distribuição normal.

3. Resultados - Pt2

O que é um Tweet

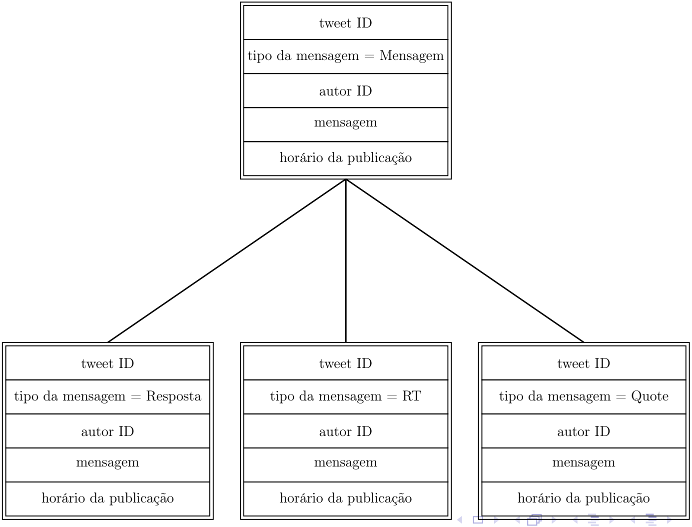
1. Um tweet pode ser abstraído como o elemento da Figura abaixo

tweet ID
tipo da mensagem
autor ID
mensagem
horário da publicação

2. **Tipo da mensagem** pode ser igual a "mensagem", "RT", "resposta", ou "Quote"

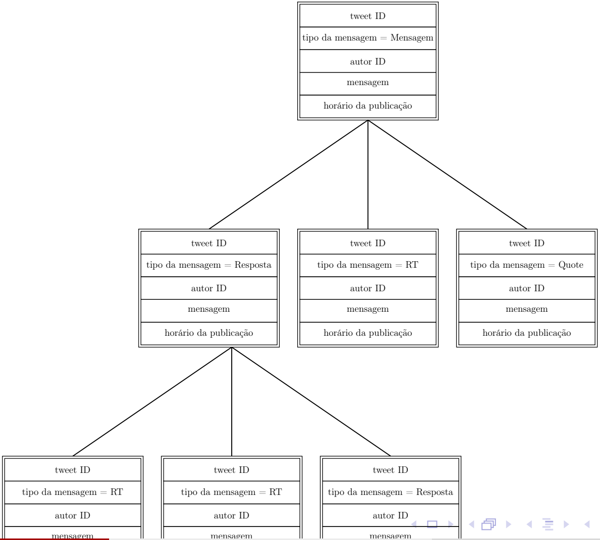
Intereção entre Tweets

1. Tweets podem interagir entre si, por exemplo:



Intereção entre Tweets

1. A intereçaão entre os tweets pode continuar indefinitivamente.



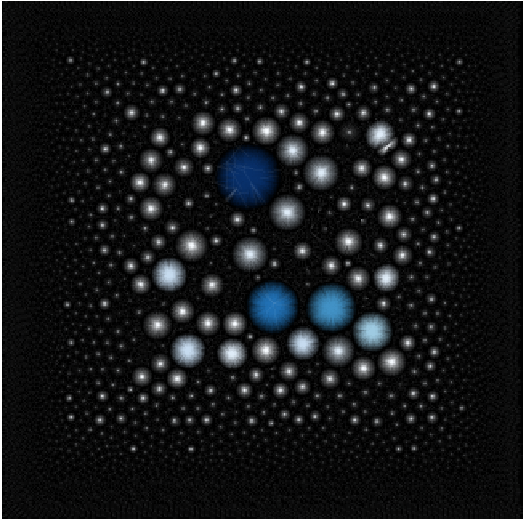
Grafo formado pelos Tweets

1. Os tweets assim nos fornecem uma sequência de árvores, que será o objeto de nossa análise;
2. **Obs**: Os IDs dos tweets são únicos;
3. **Obs**: A intereção entre os autores dos tweets ainda não está sendo levada em conta
4. **Obs** Por interação entre os autores me refiro a situação de, por exemplo, uma pessoa X seguir ou não uma pessoa Y. Sendo que X e Y publicaram um tweet

Dados coletados até agora

1. Os tweets coletados são referentes a mensagems publicadas no dia 25/03/2021 ocorrendo no período das 18h até as 21h (UTC -3) e contendo alguma das seguintes palavras chaves (em pt):
    1. vacina *ou*
    2. cloroquina *ou*
    3. covid *ou* corona *ou* covid-19 *ou*
    4. tratamento antecipado *ou* tratamento precoce *ou*
    5. azitromicina *ou*
    6. lockdown
2. **Obs** Vale notar que toda mensagem publicada no horário descrito pode ter um "parent tweet", ou seja, a mensagem coletada é uma resposta ou um RT ou uma Quote de outro tweet. Nestes casos, o tweet pai será incluso em minha análise, mesmo o mesmo tendo sido publicado fora do horário ou não contendo alguma das palavras chaves acima.

Visualização do grafo obtido



Algumas mensagens coletadas

- A seguir listo as mensagens que possuim mais de 1000 filhos, i.e., mais de 1000 pessoas interagem com estas mensagens. No total são apenas 7 mensagens.

Algumas mensagens coletadas (mensagens casuais)

1. @leelecarvalho_ "eu estou fazendo minha parte nesse lockdown, mas não milito em cima de qm ta saindo, pq não sair agora, não anula o fato de eu ter saído antes, tem uns aqui nesse tt mt hipocrita, furou a quarentena toda e agr no lockdown quer pagar de alecrim dourado kkkkk"
2. @Jouberth19 "lockdown e feriado de 10 dias pra vocês, pq eu vou continuar trabalhando normalmente"
3. @daycrvg10 "o dia q anunciarem q n há mais covid vai ser O DIA"
4. @isa_abrantes10 "Bota a Gabi da FGV nesse governo pra ver se ela não consegue vacina pra todo mundo em uma semana"

Algumas mensagens coletadas (fakenews)

1. @bicmuller "vcs tem noção que a Austrália, ZEROU as restrições pra covid ?? Sem mascara, bar aberto, estadio de futebol com 100% de capacidade CABOU COVID NA AUSTRALIA"
2. @IsabelasemZ "URGENTE - EMPRESÁRIOS ANUNICAM DOAÇÃO DE 10 MILHÕES DE VACINAS. Após reunião com Paulo Guedes, os empresários Luciano Hang e Carlos Wizard anunciaram a doação para o SUS de 10 milhões de doses da vacina contra a Covid.SERÁ QUE A IMPRENSA VAI DIZER QUE ELES SÃO BOLSONARISTAS?"
3. @BrazilFight "URGENTE: Após reunião com Min. Paulo Guedes, os empresários Luciano Hang e Carlos Wizard anunciaram a doação para o SUS de 10 milhões de doses da vacina contra a Covid. #VacinaBrasil https://t.co/pAlIXiO6Q6"

Resultados Preliminares

- Do grafo obtido iremos analisar seu diagrama de persistência de dimensão 0 e 1
- Aqui a persistência será obtida pela técnica envolvendo a homologia de caminhos persistentes (path persistent homology)
- Vale notar que para o cálculo desta homologia precisamos ter estabelecido uma matriz de peso para as arestas
- Esta matriz de peso eu adotarei como o intervalo de tempo levado para uma mensagem filha aparecer, em relação ao "parent tweet"
- Com isto obtemos