

T-MAPS: Modelo de Descrição do Cenário de Trânsito Baseado no Twitter

Bruno P. Santos
Paulo H. L. Rettore
Heitor S. Ramos
Luiz F. M. Vieira
Antonio A. F. Loureiro



Conteúdo

Contextualização,
Motivação e
Contribuições

Imprecisão, viés dos dados,
inconsistências, atribuição
espaço-temporal.

Outros trabalhos e
Próximos passos

Introdução

Coleta dos
dados

Aspectos
dos dados

Modelagem
e Avaliação

Conclusão

Fontes de dados, período,
cobertura espaço-temporal,
correlação com HERE Maps

T-MAPS, serviço de direções
e sentimento de rotas

Introdução

- ◎ **A qualidade de vida** em uma cidade é, em parte, **reflexo** da **mobilidade** que ela oferece.
- ◎ A compreensão da **mobilidade no trânsito**, tem despertado o **interesse** dos **governos** e da sociedade **acadêmica** e **empresarial**.

Introdução

◎ **A obtenção de acesso a dados é fundamental para compreender o cenário de trânsito.**

- Loops indutivos (velocidade, densidade e fluxo)
- Câmeras de trânsito
- Traces e matrizes de origem e destino

Introdução

◎ **A obtenção de acesso a dados é fundamental para compreender o cenário de trânsito.**

- Loops indutivos (velocidade, densidade e fluxo)
- Câmeras de trânsito
- Traces e matrizes de origem e destino

✘ O acesso a estes dados apresenta um grande desafio, pois em sua maioria são controlados por entidades privadas ou governamentais

Introdução

- ◎ Uma alternativa de **baixo custo para obtenção de dados** são as Mídias Sociais Baseadas em Localização (**LBSM**)
 - Ex: *Twitter e Foursquare*

Introdução

- ◎ Neste trabalho, estudamos como dados obtidos do Twitter se **relaciona** com o cenário real do trânsito
 - **Caracterização de dados do Twitter**, como fonte de dados para descrever o cenário de trânsito
 - **Desenvolvimento do T-MAPS** como um **modelo** de descrição do cenário de trânsito baseado em dados do Twitter.



XXXV
SBRC
2017

Coleta dos dados

① Usuários comuns X **Usuários especializados**



Coleta dos dados

🎯 Usuários comuns X **Usuários especializados**

Nome da conta	# tweets
@511NYC	126925
@TotalTrafficNYC	20267
@WazeTrafficNYC	7850
...	...
@NYC DOT	3680
Total de 21 contas:	655K



XXXV
SBSRC
2017



Total Traffic NYC

@TotalTrafficNYC

 Follow

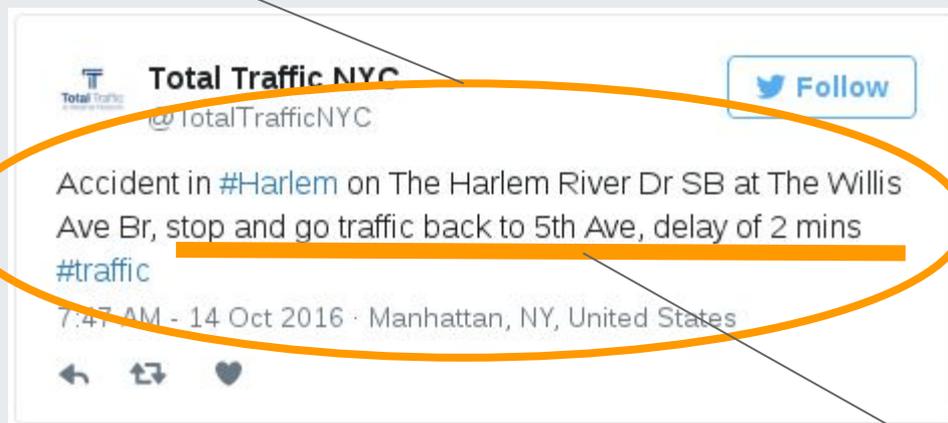
Accident in [#Harlem](#) on The Harlem River Dr SB at The Willis Ave Br, stop and go traffic back to 5th Ave, delay of 2 mins
[#traffic](#)

7:47 AM · 14 Oct 2016 · Manhattan, NY, United States





Descrição do evento.



Condições do trânsito



XXXV
SBRC
2017



Total Traffic NYC

@TotalTrafficNYC

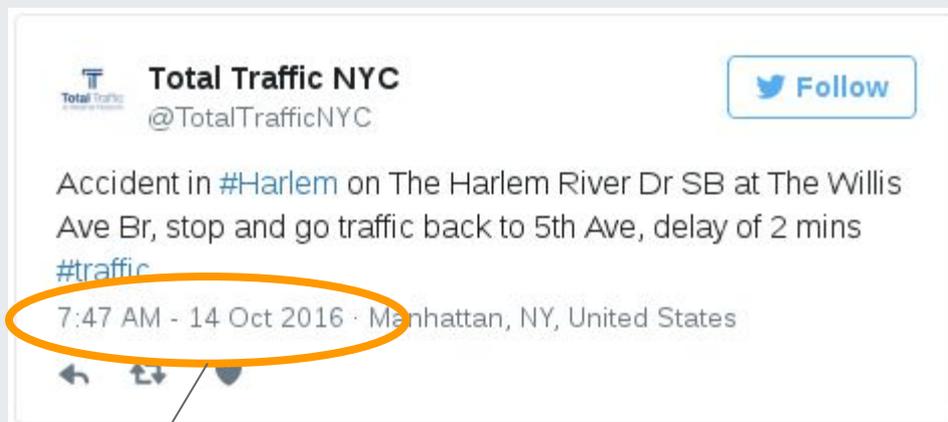
Follow

Accident in #Harlem on The Harlem River Dr SB at The Willis Ave Br, stop and go traffic back to 5th Ave, delay of 2 mins
#traffic

7:47 AM - 14 Oct 2016 **Manhattan, NY, United States**



**Descrição da localização e,
em geral, geo-referência.**



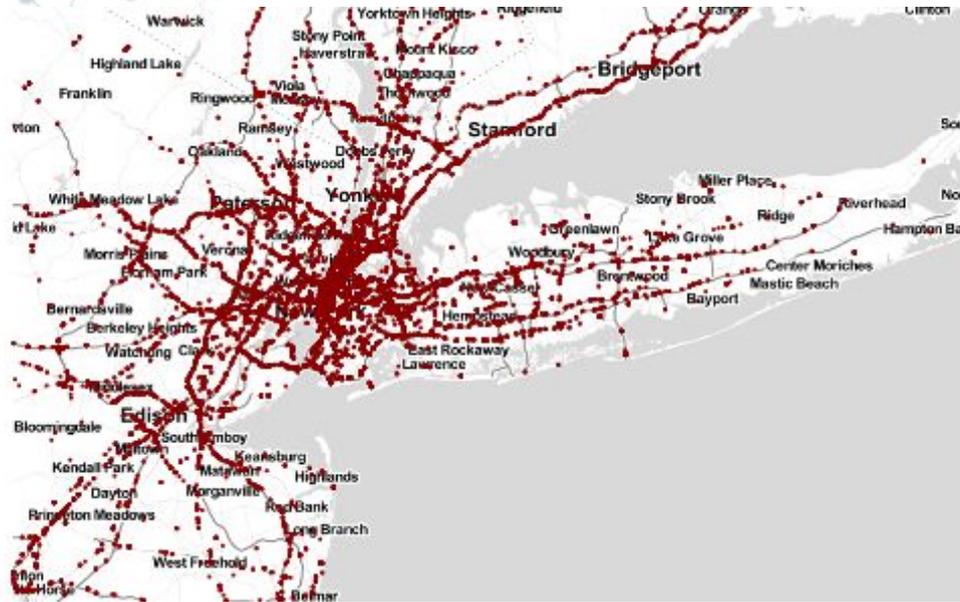
Horário do tweet.



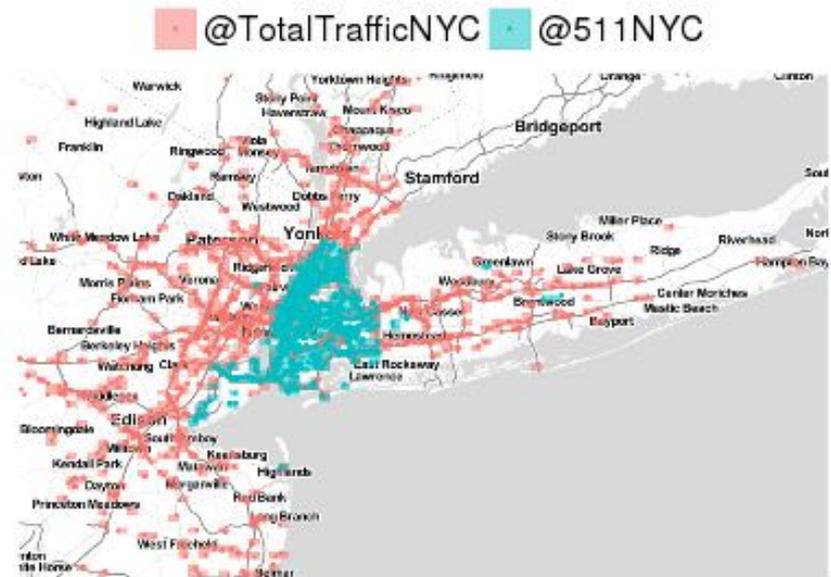
XXXV
S B R C
2017

Coleta dos dados

Cobertura Espacial



tweets em NY

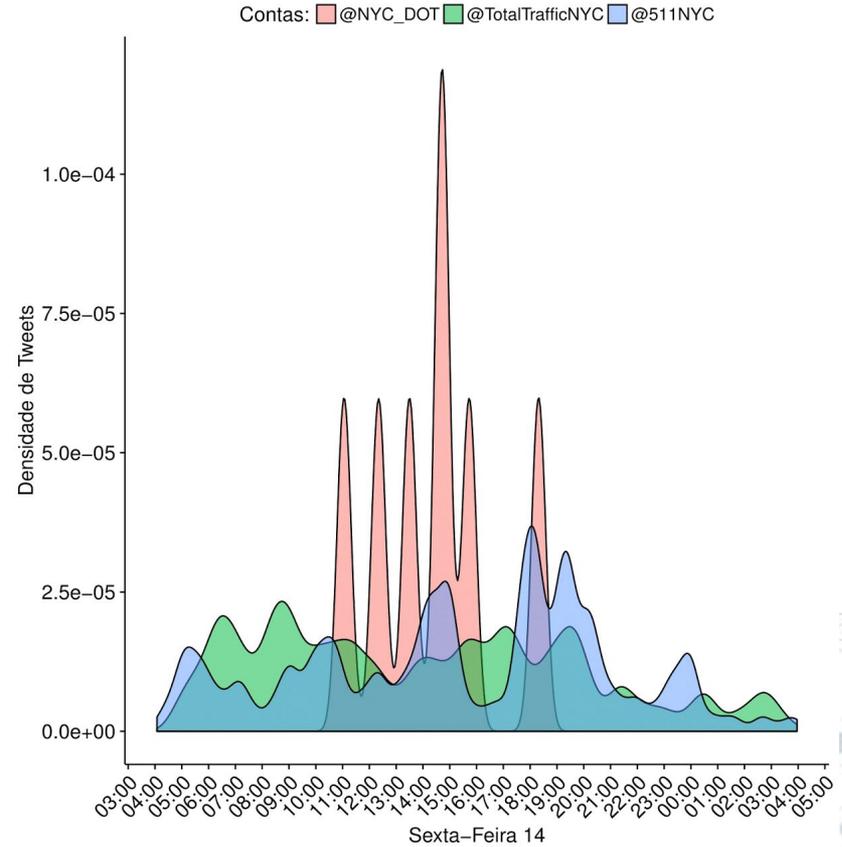
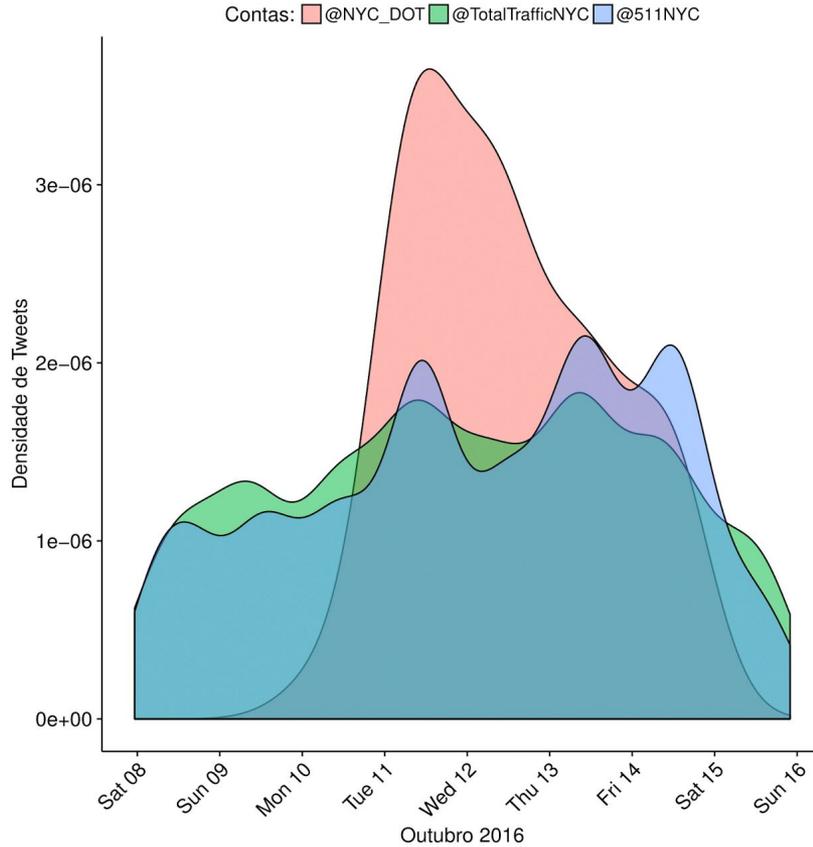


Cobertura espacial de duas contas em NY

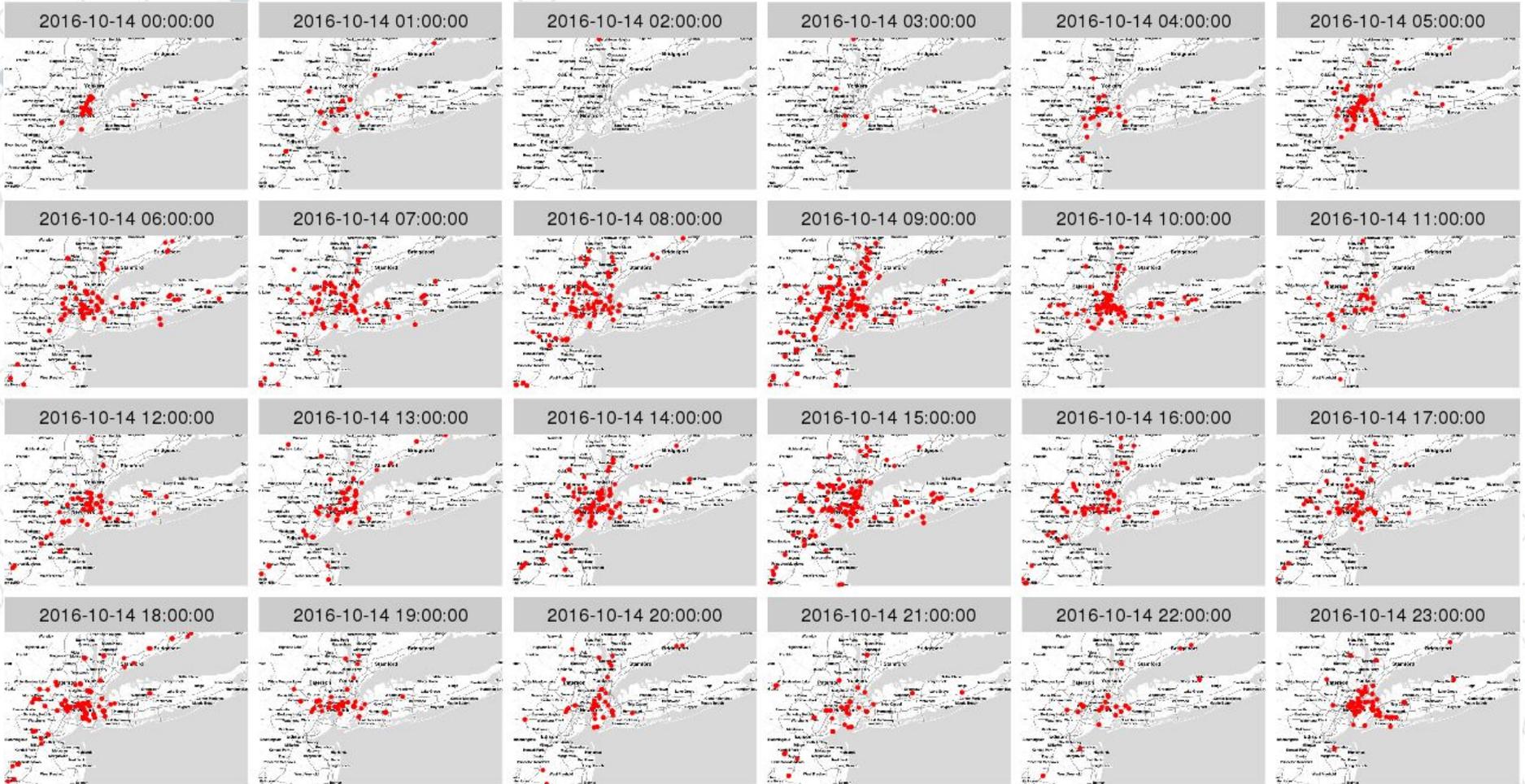


Coleta dos dados

Cobertura Temporal



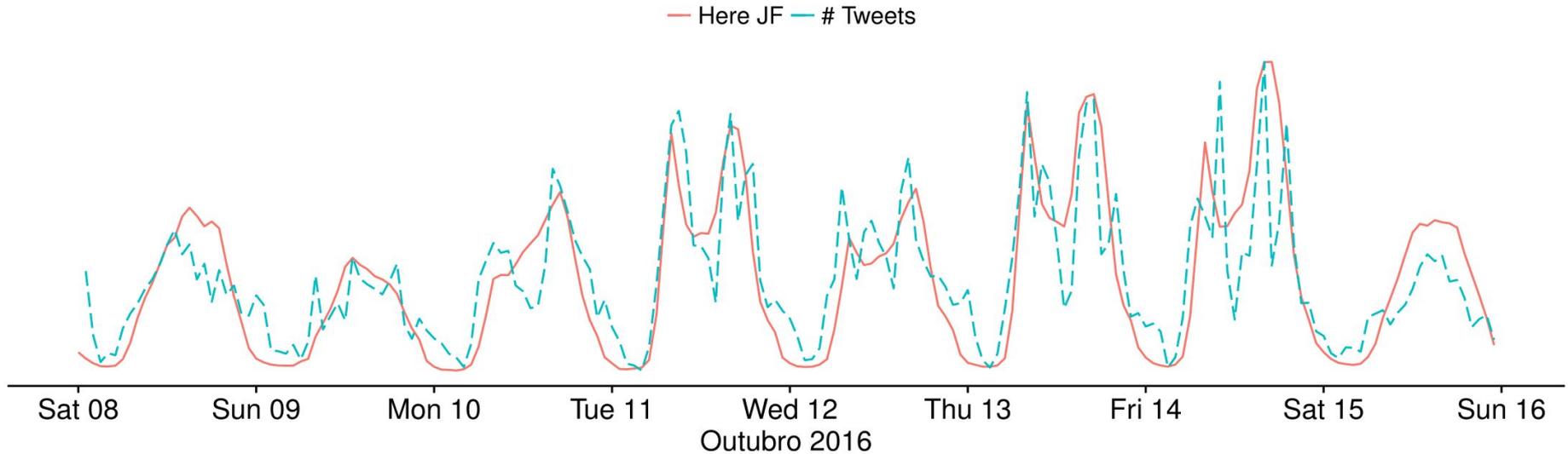
**Cobertura temporal de
três contas**





Coleta dos dados

Existe correlação entre *tweets* e trânsito?



Dados:

- Jam Factor (Here developer API)
- # contagem simples dos *tweets*



Existe correlação entre tweets e trânsito?

Correlações

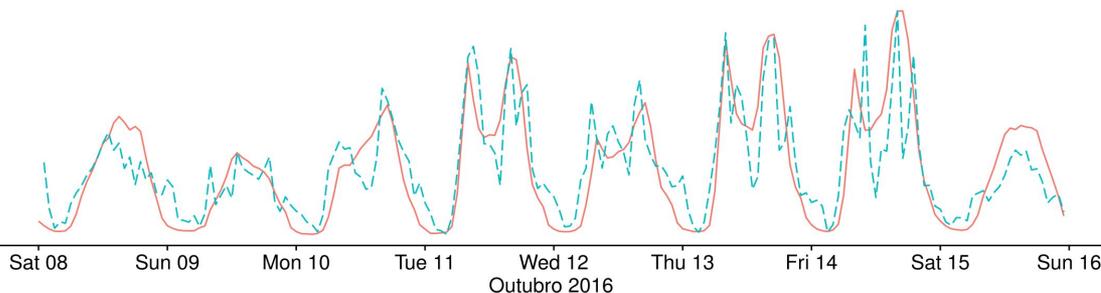
- Pearson vs **Spearman**

Coeficiente de correlação $\rho = 0.81$

Interpretação

- Indica que existe alta correlação
- Por ser positiva
 - Em momentos que o fator de congestionamento cresce os *tweets* também crescem

— Here JF — # Tweets





XXXV
SBRC
2017

Aspectos dos dados



Imprecisão

Dados podem ser incompletos, vagos ou ter níveis granularidade



Viés dos usuários

Usuários comuns vs usuários especialistas. Interesses próprios.



Inconsistências

Dados conflitantes ou fora de ordem



Atribuição espaço-temporal

Sem geolocalização, o dado diz respeito ao futuro ou passado?

Aspectos dos dados

Imprecisão - Incompleto

“



*Agora 8:00AM um acidente na Av. Antônio Carlos
#BH #trafegoRuim #asustado*

Imprecisão - Incompleto

“



*Agora 8:00AM um acidente na Av. Antônio Carlos
#BH #trafeqoRuim #asustado*

- Quando?
- Qual é o evento?
- Condições do trânsito?
- Sentimento?

Imprecisão - Incompleto

“



*Agora 8:00AM um acidente na Av. Antônio Carlos
#BH #tráfegoRuim #asustado*

- Quando?
- Qual é o evento?
- Sentimento?
- Condições do trânsito?

- ✗ **Geolocalização**
- ✗ **Descrição textual incompleta**

Imprecisão - Vago

“



*Agora 8:00AM um acidente na Av. Antônio Carlos
#BH #trafeqoRuim #asustado*

- Quando?
- Qual é o evento?
- Sentimento?
- Condições do trânsito?

- ✗ **Qual ponto da Av.?**
- ✗ **tweet limitado (150 carac.)**

Imprecisão - Granularidade

“



 **Total Traffic NYC**
@TotalTrafficNYC 

Accident in #Harlem on The Harlem River Dr SB at The Willis Ave Br, stop and go traffic back to 5th Ave, delay of 2 mins
#traffic

7:47 AM - 14 Oct 2016 · Manhattan, NY, United States

 **OficialBHTRANS** 
@OficialBHTRANS 

12h47 Área Hospitalar: Trânsito intenso nos dois sentidos da Alfredo Balena.

12:47 PM - 2 May 2017

  1  2

↓ **Granularidade**
Apresentam informações suficientes para descrever precisamente:

- **Local**
- **Sentido**
- **gravidade, etc..**

↑ **Granularidade**
Apresentam informações suficientes para descrever uma visão macro dos trânsito

Aspectos dos dados

Viés dos usuários

- ◎ Visão de um engarrafamento
 - Usuário de uma metrópole X de um vilarejo
- ◎ Foco de informação
 - Contas de usuários especialistas
 - Ex: jornais e FM

- **O trânsito é intenso para um e normal para outro ?**
- **Informação direcionada**



XXXV
SBRC
2017



Aspectos dos dados

Atribuição espacial e temporal

- ◎ Ponto crítico dos aspectos dos dados
- ◎ Dados sem tais atributos podem não fazer sentido na construção do cenário de trânsito

Aspectos dos dados

Atribuição espacial e temporal

“



Atribuir um tweet a um ponto do tempo e do espaço pode não ser trivial mesmo que a informação esteja presente.



XX XV
S B R C
2 0 1 7



Aspectos dos dados

Atribuição espacial e temporal



Um *tweet* geralmente é

- Desestruturado, Limitado (impreciso, tamanho, coerência)
- Subjetivo:
 - R significa Rua ou Rodovia?
- Possui um *timestamp*
 - Faz sentido para o momento do evento?
 - Qual é a validade de uma publicação?



Aspectos dos dados

Atribuição espacial e temporal



Named Entity Recognition:

	<u>Loc</u>	<u>Location</u>	<u>Location</u>	<u>ORDINAL</u>	
1	Accident in Harlem on The Harlem River Dr SB at the Willis Ave Br, stop and go traffic back to 5th Ave, delay of				
	<u>Dur</u>				
	2 min traffic				

<http://nlp.stanford.edu:8080/corenlp/>



Modelagem e avaliação do Twitter Maps

① Processo de modelagem

1. Aquisição de informação
2. Filtragem e fusão de dados
3. Métricas de custo

② Avaliação

1. Serviço de rotas
2. Serviço de sentimento das regiões



Modelagem e avaliação do Twitter Maps

Passo 1 - Aquisição de informações



Cidade de Nova Iorque



Dados de plataformas LBSM



Modelagem e avaliação do Twitter Maps

Passo 2 - Filtragem e fusão dos dados



$G(V,A)$

- $V(G)$ são as divisões da região
- $A(G)$ são arestas em ligam regiões adjacentes

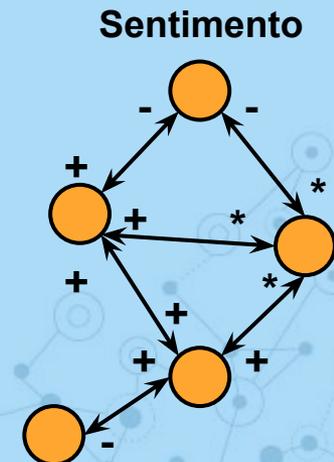
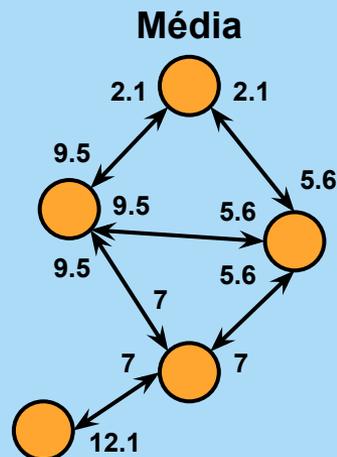
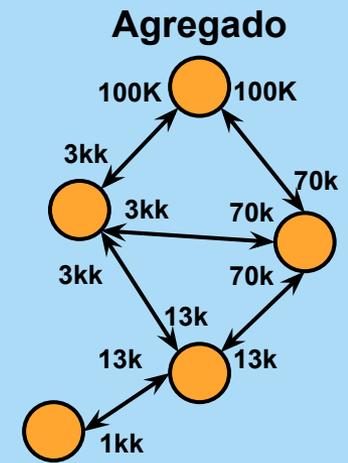
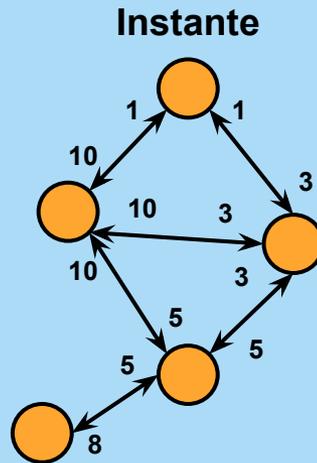


- Atribuição espacial, temporal
- Filtragem para obter dados da região
- Remoção de dados inconsistentes

Modelagem e avaliação do Twitter Maps

Passo 3 - Fusão dos dados e métricas de custo

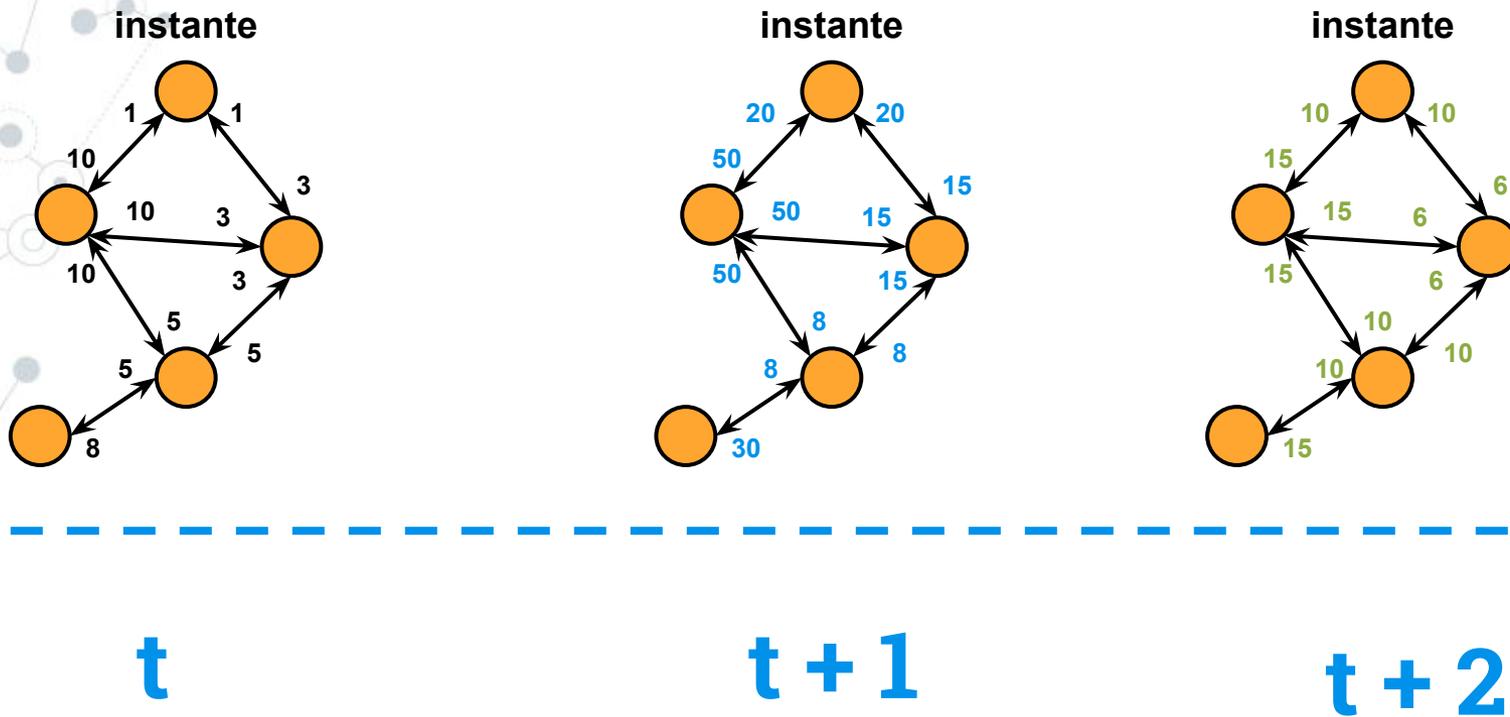
(Uma descrição do cenário de trânsito)





Modelagem e avaliação do Twitter Maps

discretização do tempo - Métrica - Instantâneo





Modelagem e avaliação do Twitter Maps

Avaliação



Manhattan (região de interesse)

- 29 divisões (unidades administrativas)
- 21 contas especialistas do Twitter.
 - ~ 280 K tweets (geo-localizados)
 - Outubro - Dezembro de 2016



T-MAPS

- Aplicação do algoritmo de Dijkstra

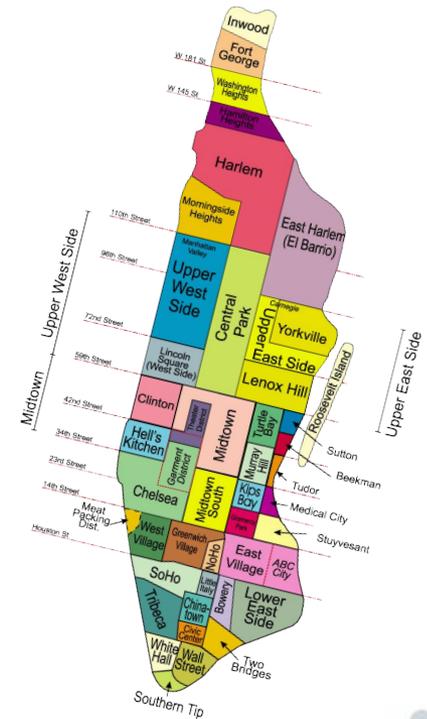


Google Directions¹ (usado como representação fiel do cenário de trânsito)



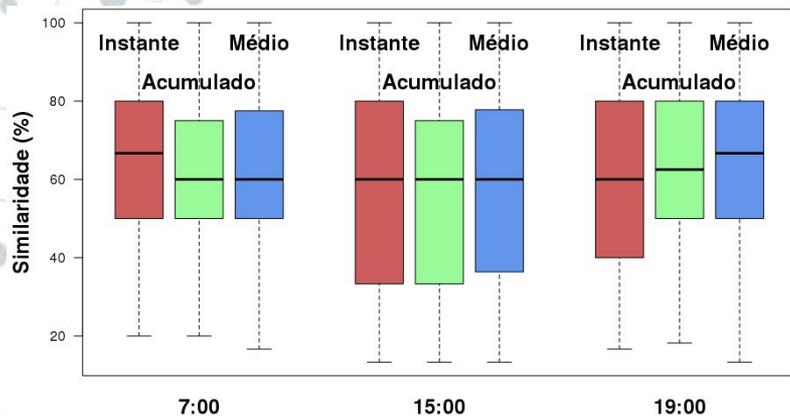
Similaridade

- É o percentual de interseção das divisões (bairros) recomendadas pelo T-MAPS e Google Directions

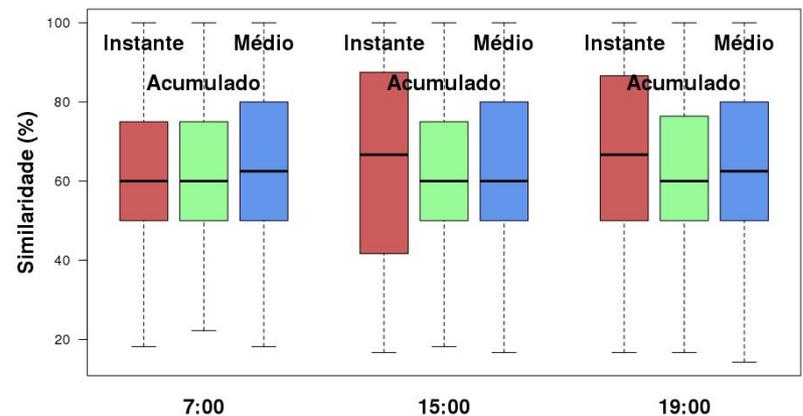


Modelagem e avaliação do Twitter Maps

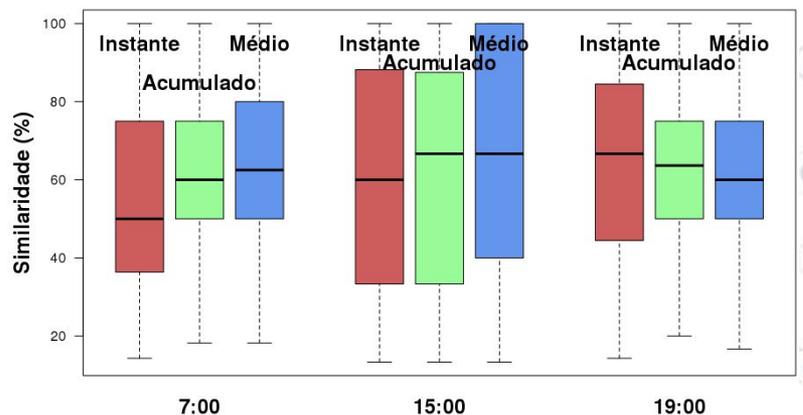
Avaliação



7 de Dezembro de 2016



8 de Dezembro de 2016



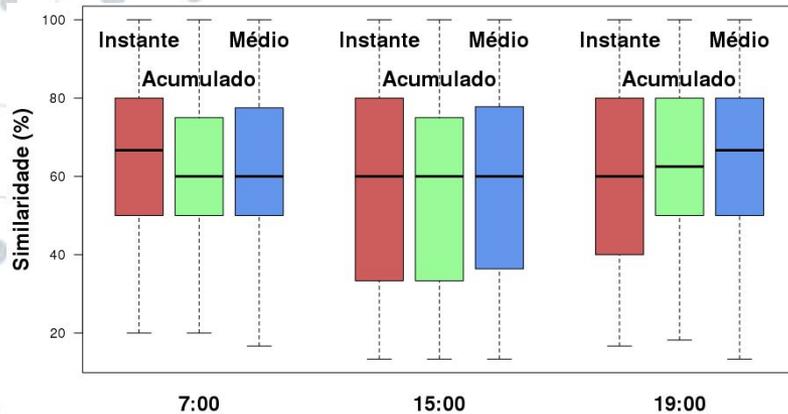
9 de Dezembro de 2016

Rotas usando a métrica **INSTANTE** apresentam maior variação de similaridade.

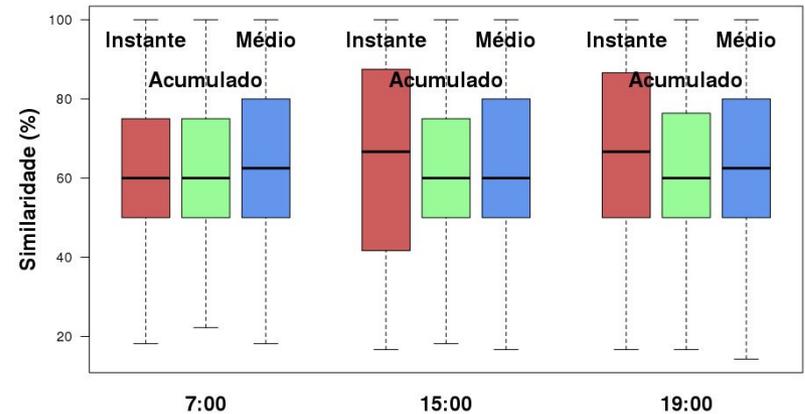
- A **mediana varia** entre **50% a 60%**

Modelagem e avaliação do Twitter Maps

Avaliação



7 de Dezembro de 2016



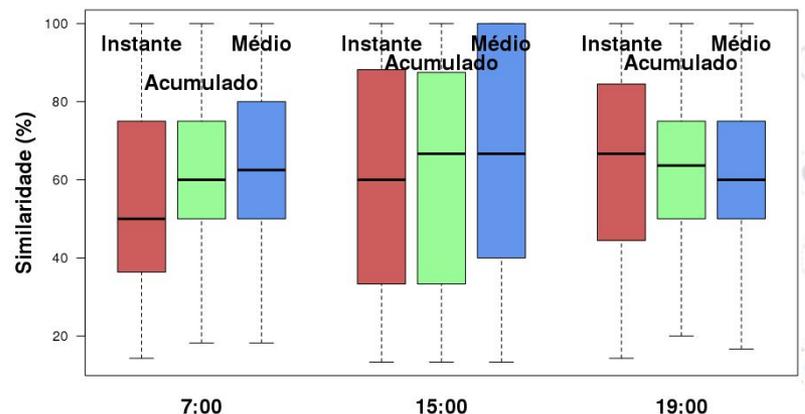
8 de Dezembro de 2016

Rotas usando a métrica **ACUMULADO** apresentam maior variação de similaridade.

- A **mediana varia** entre **60% a 66%**

Rotas usando a métrica **MÉDIA** apresentam maior variação de similaridade.

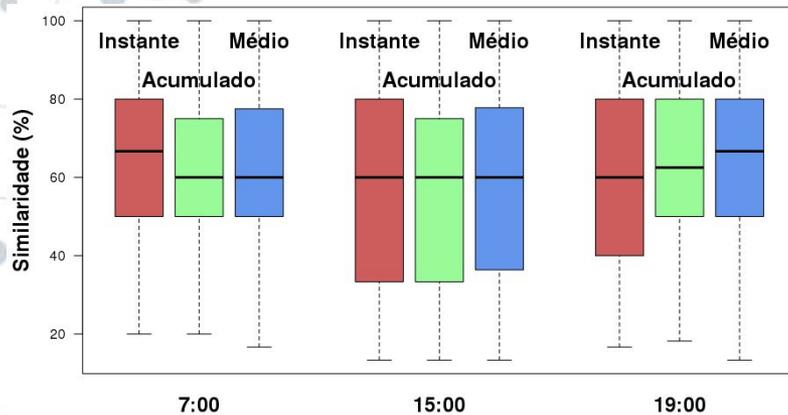
- A **mediana varia** entre **63% a 67%**



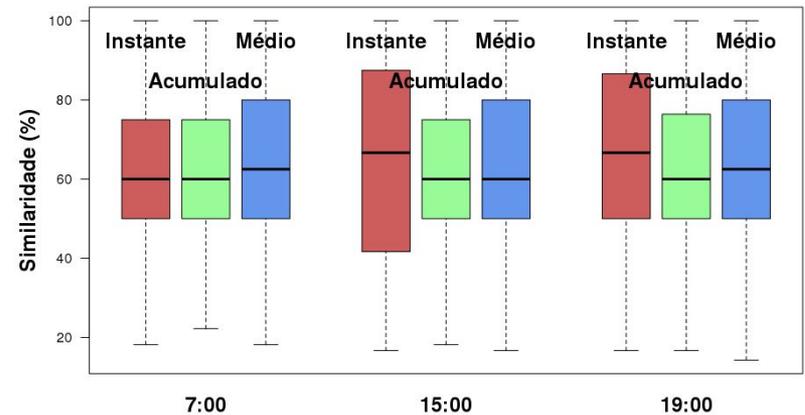
9 de Dezembro de 2016

Modelagem e avaliação do Twitter Maps

Avaliação

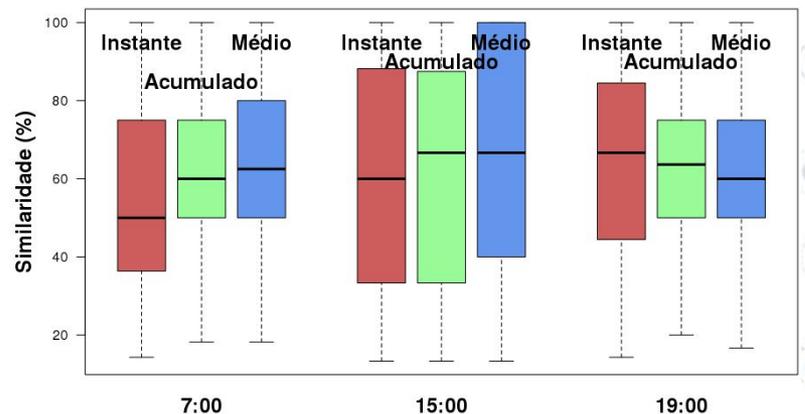


7 de Dezembro de 2016



8 de Dezembro de 2016

- Em média, metade das rotas, apresentam 62% de similaridade
- 25% das rotas apresentam similaridade entre 87% e 100%



9 de Dezembro de 2016

Modelagem e avaliação do T-Maps

Avaliação

 Coleta dos *tweets*

 RM pontuação

 RM palavras de parada
(*stop words*)

✗ ‘/,.,:*”?!/...

✗ the, is, at, which...

Bibliotecas do R:
[syuzhet, tm, stringr, wordcloud]

Modelagem e avaliação do T-Maps

Avaliação

- Coleta dos *tweets*
- RM pontuação
- RM palavras de parada (*stop words*)
- Stemming

Jamming, jammed → **Jam**
Ave, Av → **Avenue**
St → **Street**

Modelagem e avaliação do T-Maps

Avaliação

 Coleta dos tweets

 RM pontuação

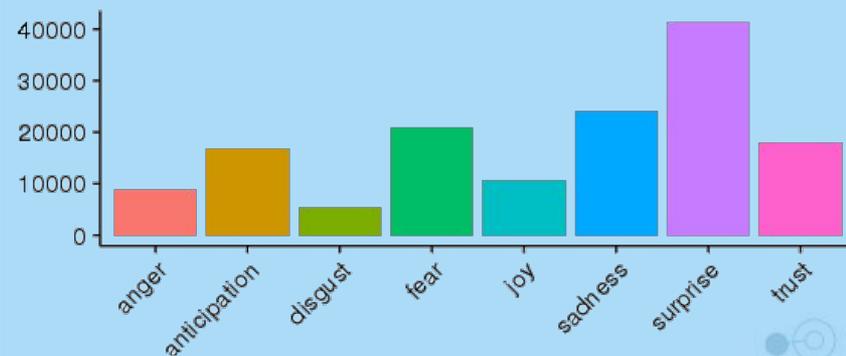
 RM palavras de parada
(*stop words*)

 Stemming



Score

Nuvem de
palavras



Modelagem e avaliação do T-Maps

Avaliação

 **Coleta dos tweets**

 RM pontuação

 RM palavras de parada
(*stop words*)

 **Stemming**



 **Score**



 **Pos ou Neg?**

 **Nuvem de palavras**

##	Ang.	Anticip.	Disg.	Fear	Joy	Sad.	Surpri.	Trust	Neg.	Pos.
## 1	0	1	0	0	1	0	0	1	0	1
## 2	0	0	0	0	0	0	0	0	0	0
## 3	0	0	0	0	0	0	0	0	0	0
## 4	0	1	1	0	1	1	0	2	1	2
## 5	0	1	0	0	1	0	1	2	0	2
## 6	0	0	0	0	0	0	0	0	0	0

Modelagem e avaliação do T-Maps

Avaliação

 **Coleta dos tweets**

 RM pontuação

 RM palavras de parada
(*stop words*)

 **Stemming**





 **Score**



 **Pos ou Neg?**



 **Rotas e mapas**

 **Nuvem de palavras**



Conclusões

- ◎ Apresentamos um estudo de **caracterização e relacionamento** entre dados do **Twitter** e o **cenário de trânsito**
- ◎ Apresentamos o T-MAPS
 - Rotas sugeridas foram, em média, **62% similares** com as rotas do G. Directions
 - **25%** das rotas avaliadas foram obtidos graus de **similaridade entre 87% e 100%**

Conclusões

🎯 Questões em aberto

1. Como medir a confiabilidade das fontes e validade dos dados no tempo?
2. Como explorar os textos dos *tweets*, para extrair mais informações sobre os eventos?
 - Ferramentas de NLP específicas para *tweets*
3. Como saber se um *tweet* é válido?
4. Como estender o T-MAPS para regiões com maiores dimensões?
 - Problema computacional



XXV
SBRC
2017

Obrigado!

Perguntas?

bruno.ps@dcc.ufmg.br