Mineração de Dados

Aula 5 – parte 2

Especialização em Ciência de Dados e suas Aplicações

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Estudo de diferenças culturais

You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare.
ICWS'15

Estudo de diferenças culturais

Grande desafio: encontrar dados apropriados para uso

- **Métodos tradicionais:** Questionários
  - Não escalam
  - Difícil de detectar mudanças dinâmicas

Estudo de diferenças culturais

Grande desafio: encontrar dados apropriados para uso

- **Métodos tradicionais:** Questionários
  - Não escalam
  - Difícil de detectar mudanças dinâmicas

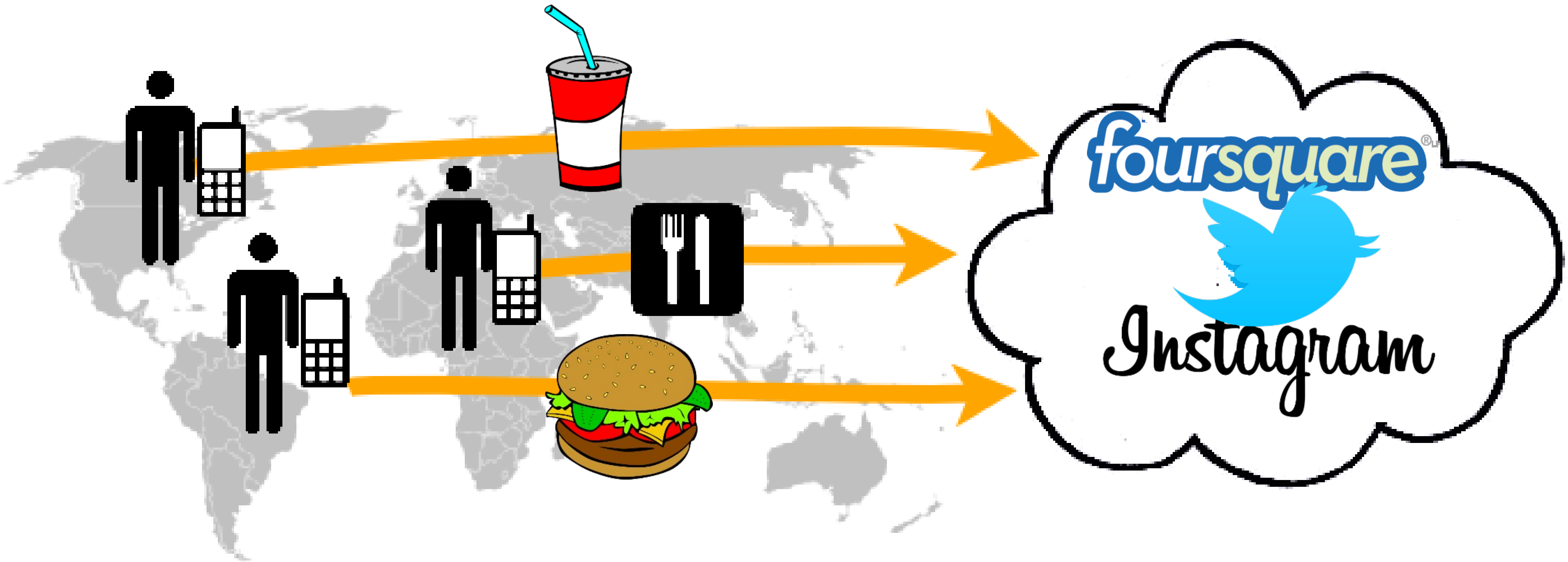
É possível propor algum método alternativo?

Você é o que você come

Hábitos alimentares e de bebida são elementos fundamentais em uma cultura

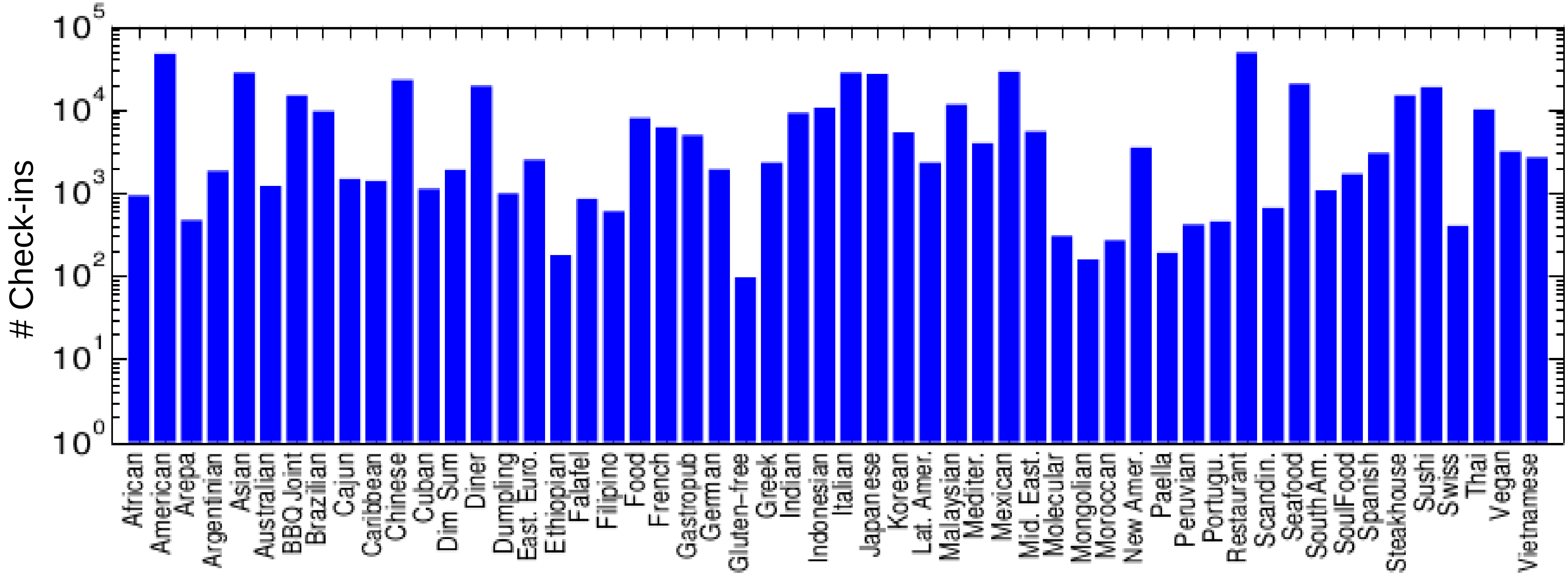
Redes de sensoriamento participativo

Sensoriamento de atividades humanas em larga escala!

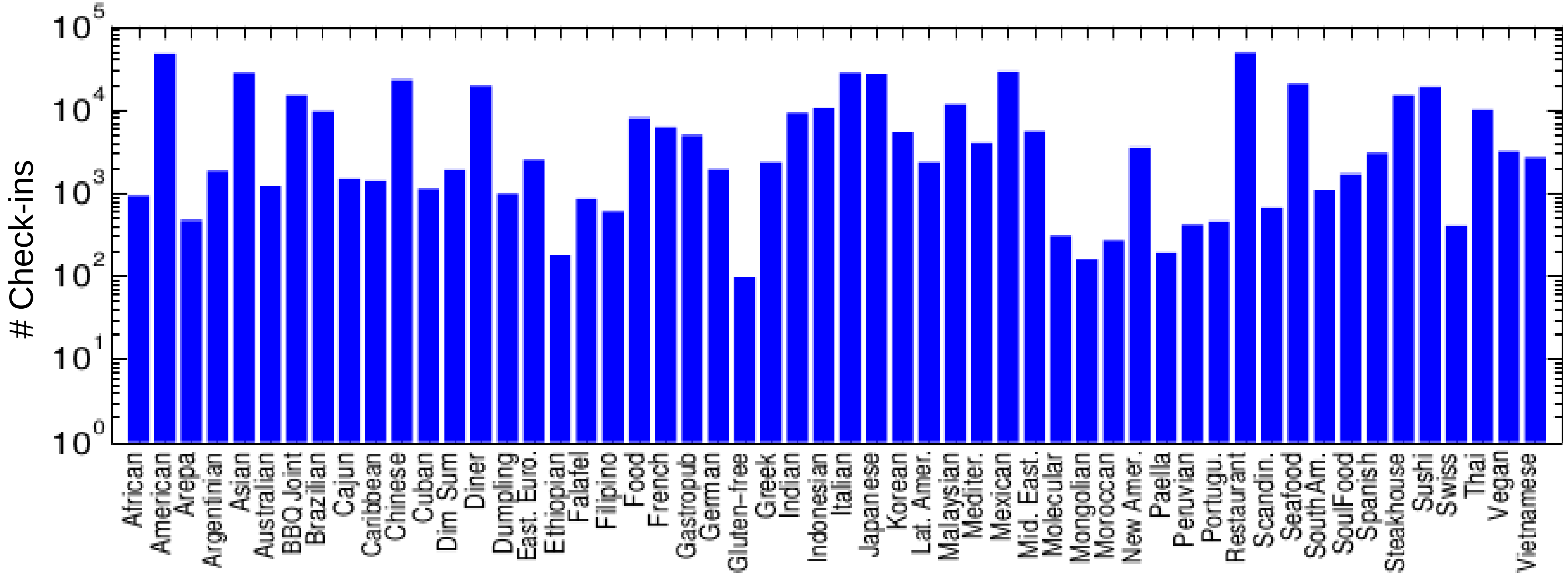


Oportunidade sem precedentes para estudar diferenças culturais em escala global e baixo custo

Categorias Slow Food

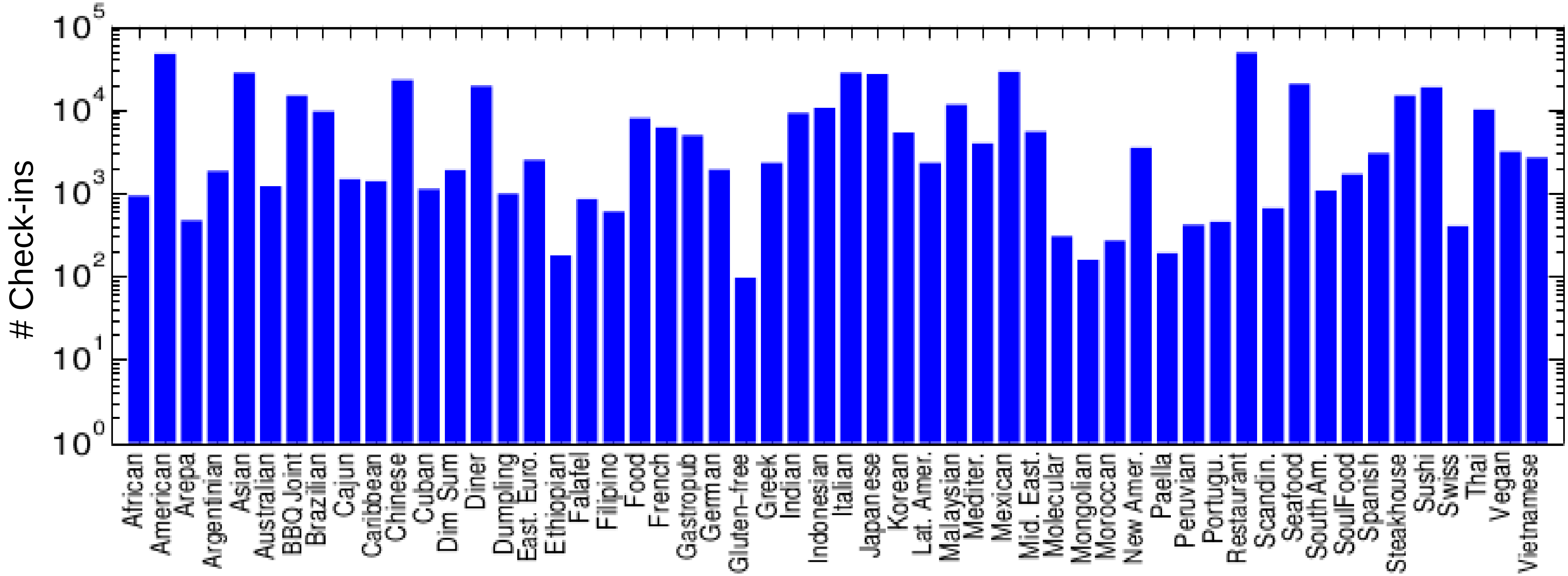


Mapeando os hábitos



Mapeia cada usuário n_i n $F_i = f_{1^i}, f_{2^i}, \ldots, f_{m^i}$

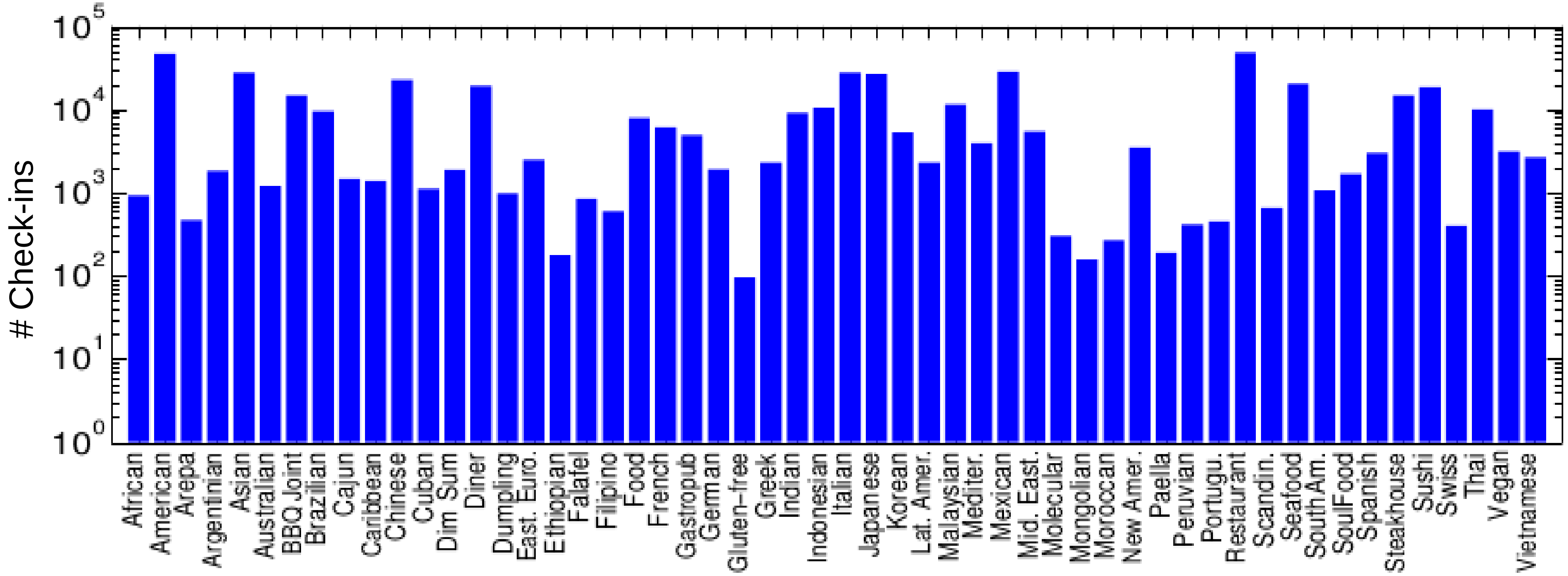
Mapeando os hábitos



Mapeia cada usuário n_i n $F_i = f_{1^i}, f_{2^i}, \ldots, f_{m^i}$

Como perguntas em um survey!

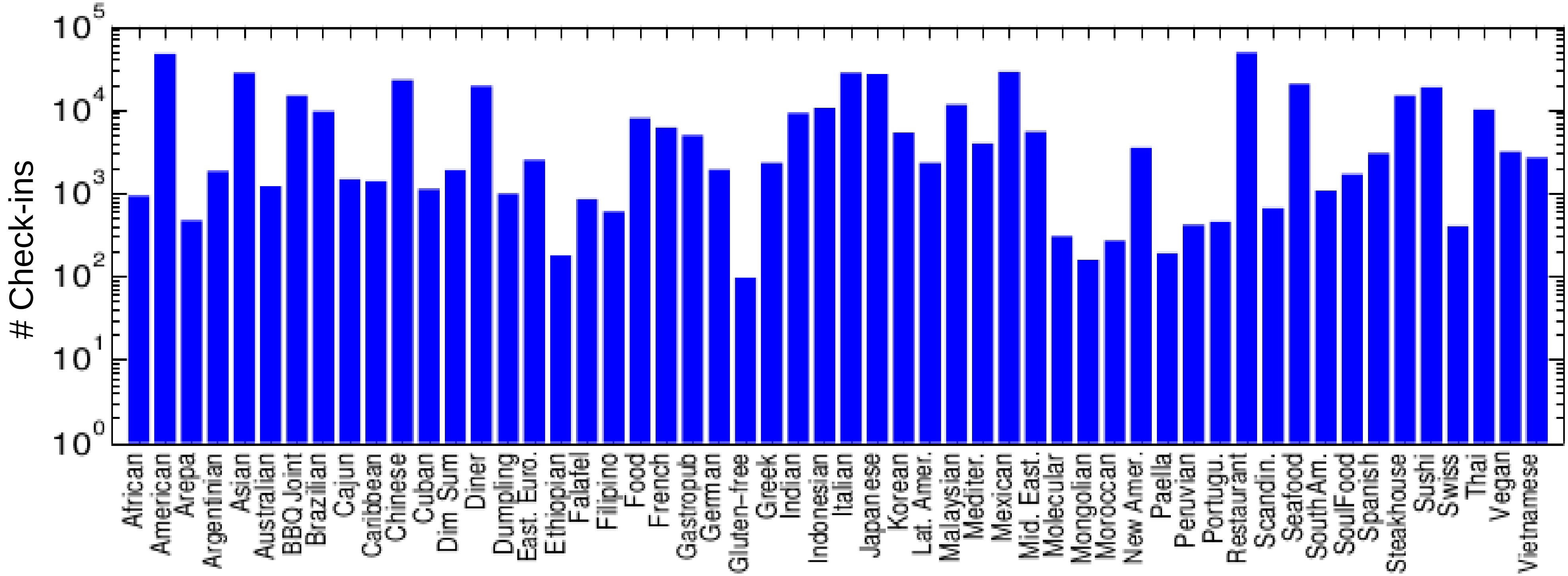
Mapeando os hábitos



Mapeia cada usuário n_i n $F_i = f_{1^i}, f_{2^i}, \ldots, f_{m^i}$

$f_{k^i} = 0|1$ representa se o usuário n_i gosta de f_k

Mapeando os hábitos



Mapeia cada usuário n_i n $F_i = f_{1^i}, f_{2^i}, \ldots, f_{m^i}$

$f_{k^i} = 0|1$ representa se o usuário n_i gosta de f_k

Respostas dos usuários

Data requirements

Data from LBSNs can be used if and only if:

Data requirements

Data from LBSNs can be used if and only if:

1 - Associate a user to its location;

Brazilian

Data requirements

Data from LBSNs can be used if and only if:

1 - Associate a user to its location;

2 - Extract finite set of preferences from the data;

Brazilian

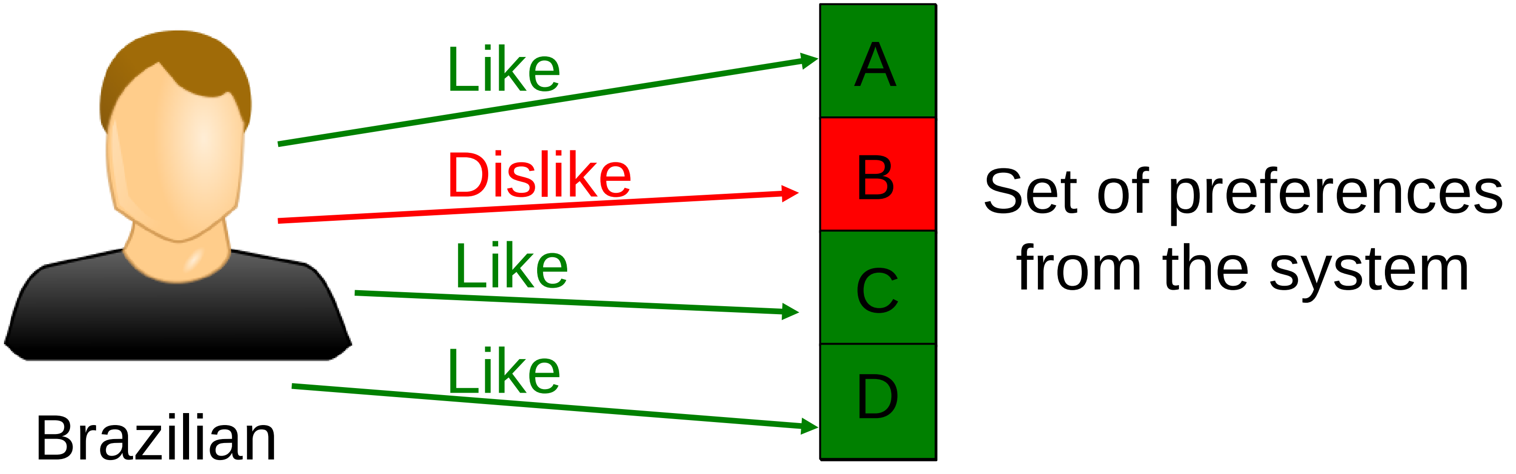
A
B
C
D

Set of preferences from the system

Data requirements

Data from LBSNs can be used if and only if:

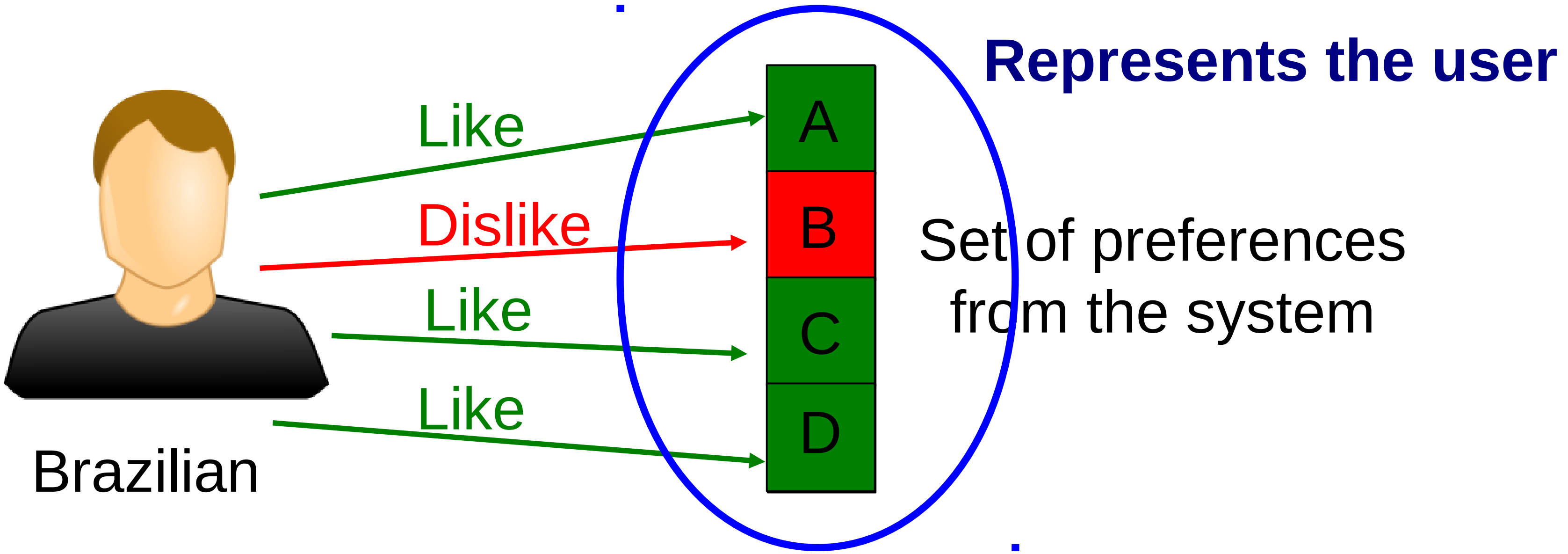
1 - Associate a user to its location;
2 - Extract finite set of preferences from the data;
3 - Map users' actions into the preferences.



Data requirements

Data from LBSNs can be used if and only if:

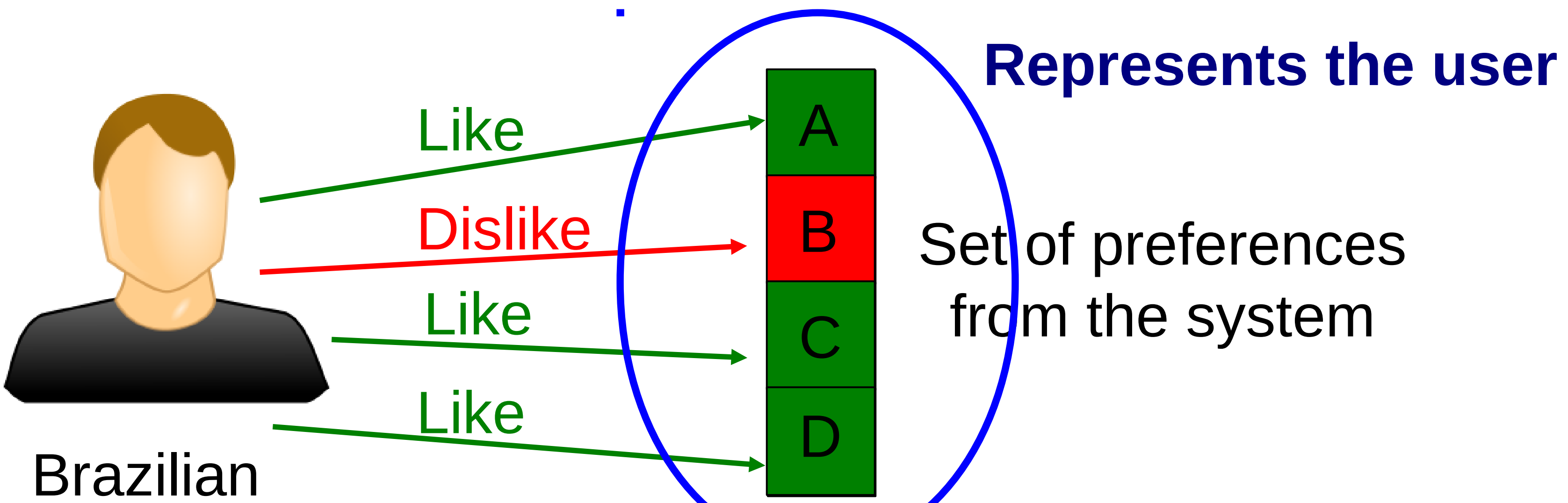
1 - Associate a user to its location;
2 - Extract finite set of preferences from the data;
3 - Map users' actions into the preferences.



Data requirements

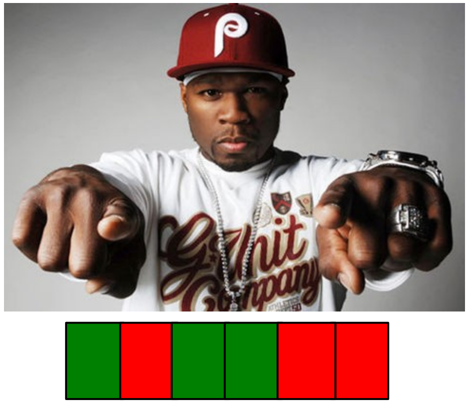
Data from LBSNs can be used if and only if:

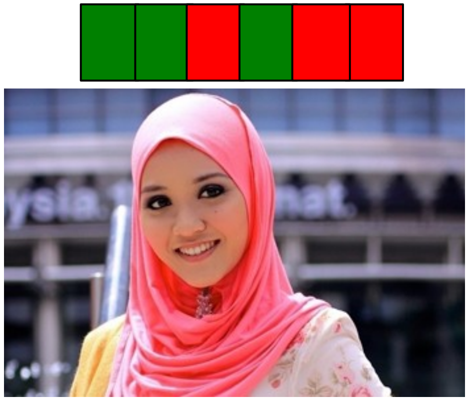
1 - Associate a user to its location;
2 - Extract finite set of preferences from the data;
3 - Map users' actions into the preferences.



We demonstrate with Foursquare data

Analise cultural de indivíduos



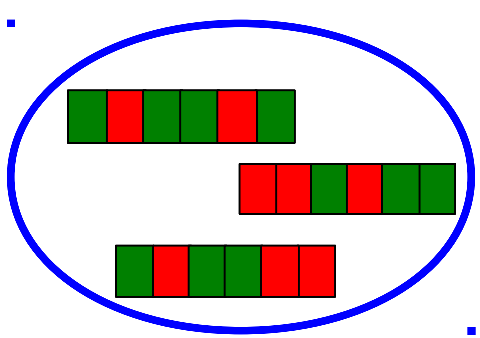


Can we define cultural signatures of different areas around the world?

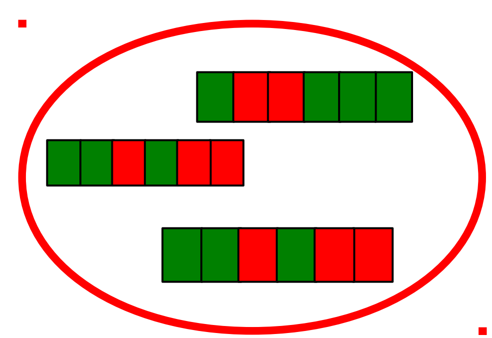
Extraction of cultural signatures

Spatial evaluation

For a given geographical area:



Area a (Belo Horizonte)



Area b (New York City)

Extraction of cultural signatures

Spatial evaluation

For a given geographical area:

- Aggregate all users' preference in normalized vectors

Area a (Belo Horizonte)

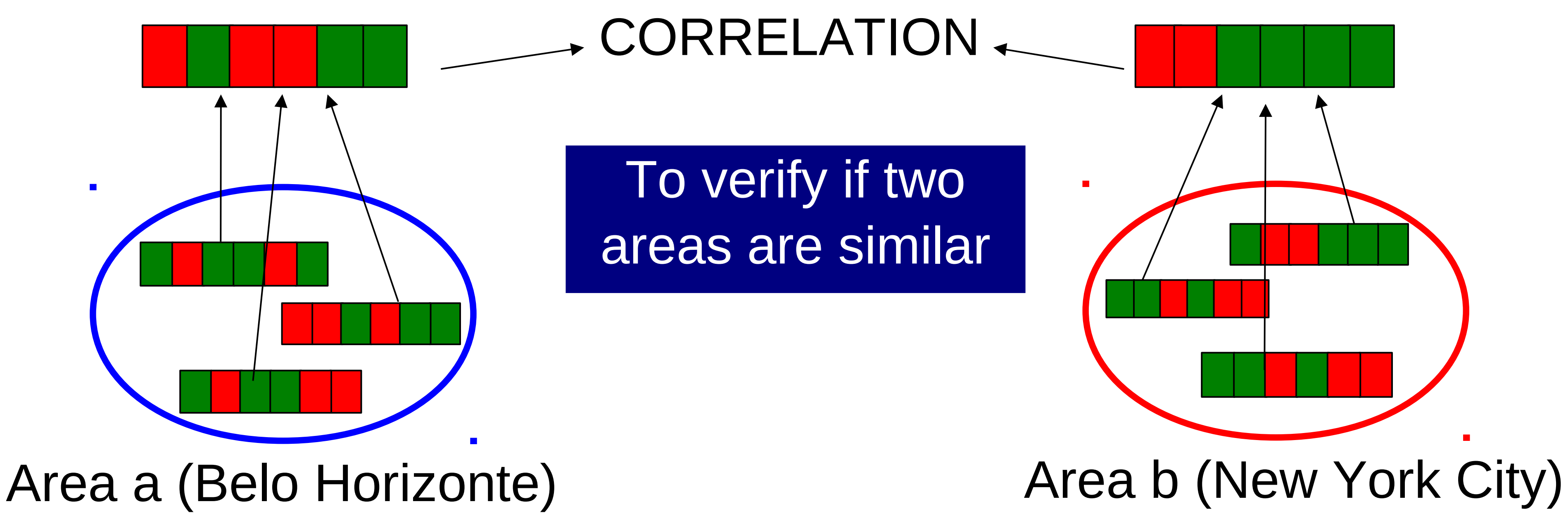
Area b (New York City)

Extraction of cultural signatures

Spatial evaluation

For a given geographical area:

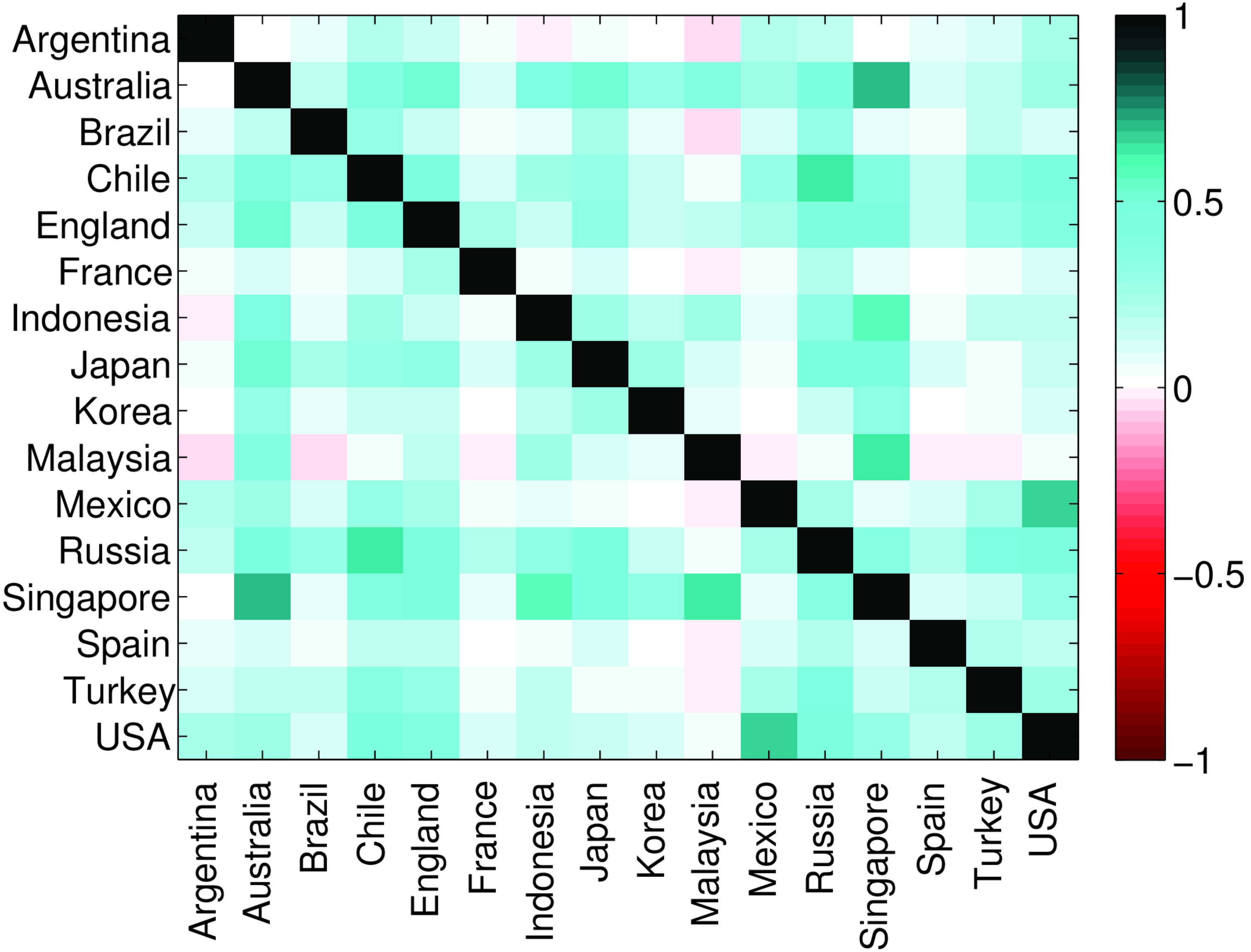
- Aggregate all users' preference in normalized vectors



Extraction of cultural signatures

Spatial evaluation

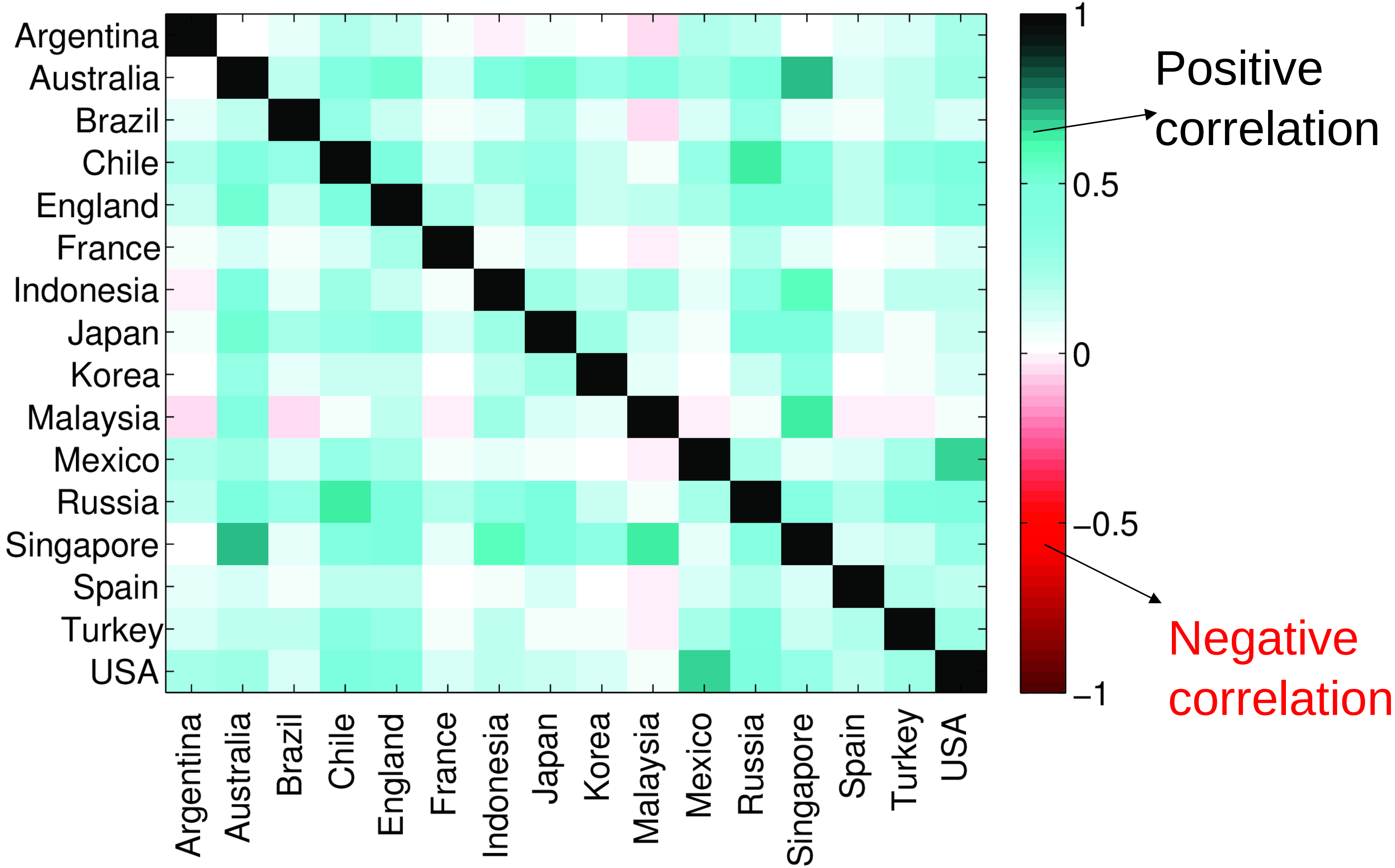
Results for **countries**



Extraction of cultural signatures

Spatial evaluation

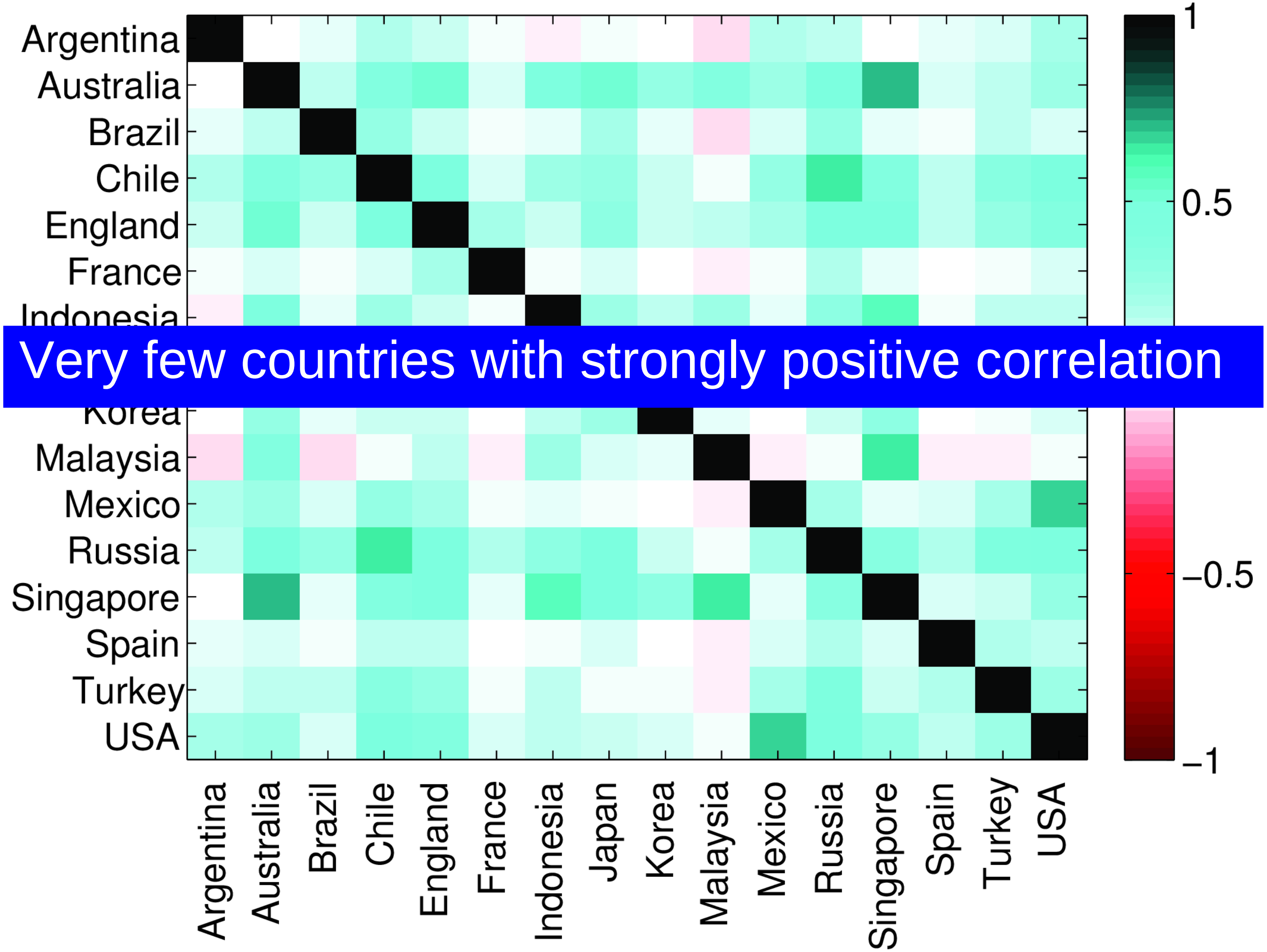
Results for **countries**



Extraction of cultural signatures

Spatial evaluation

Results for **countries**

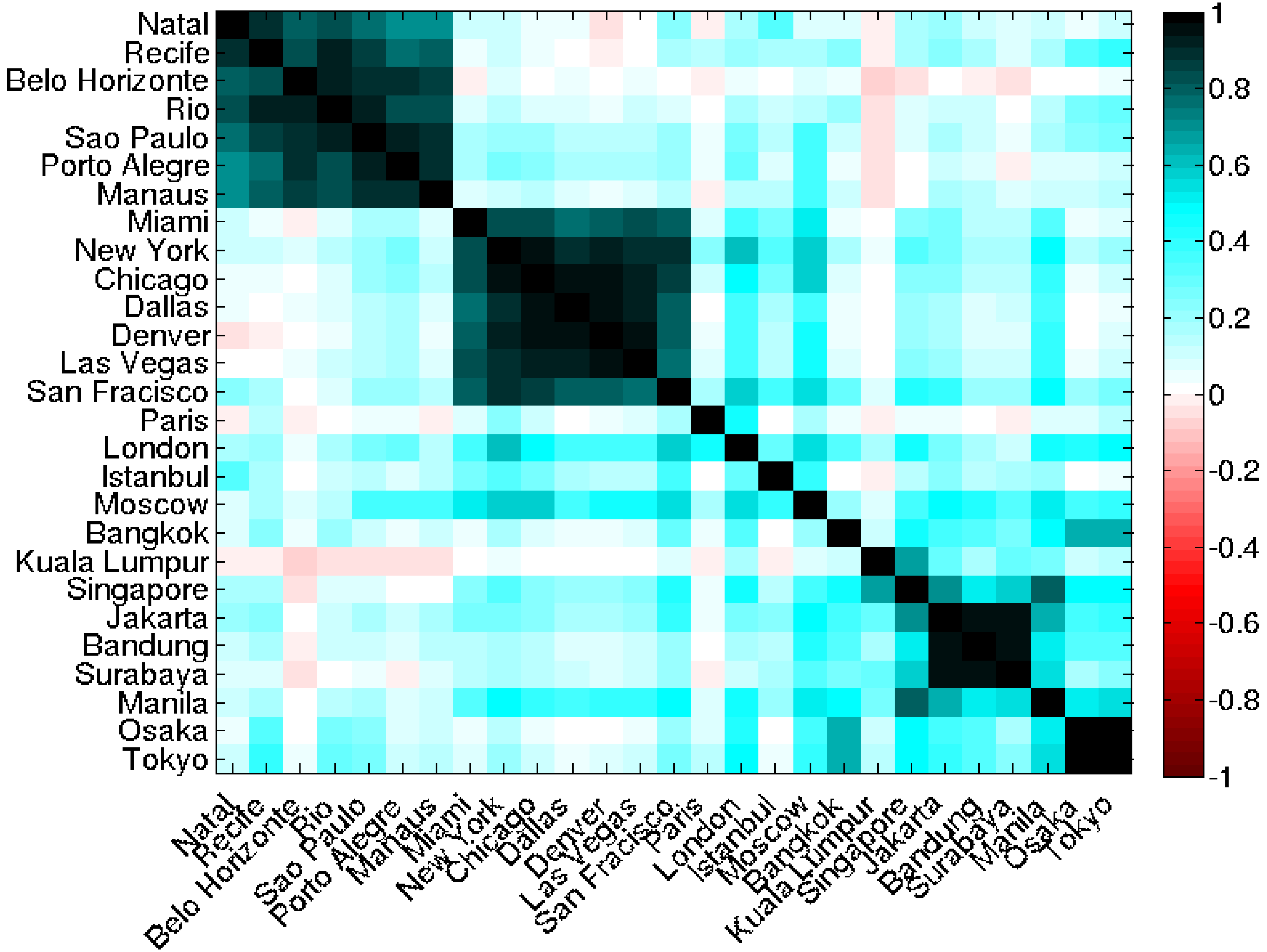


Very few countries with strongly positive correlation

Extraction of cultural signatures

Spatial evaluation

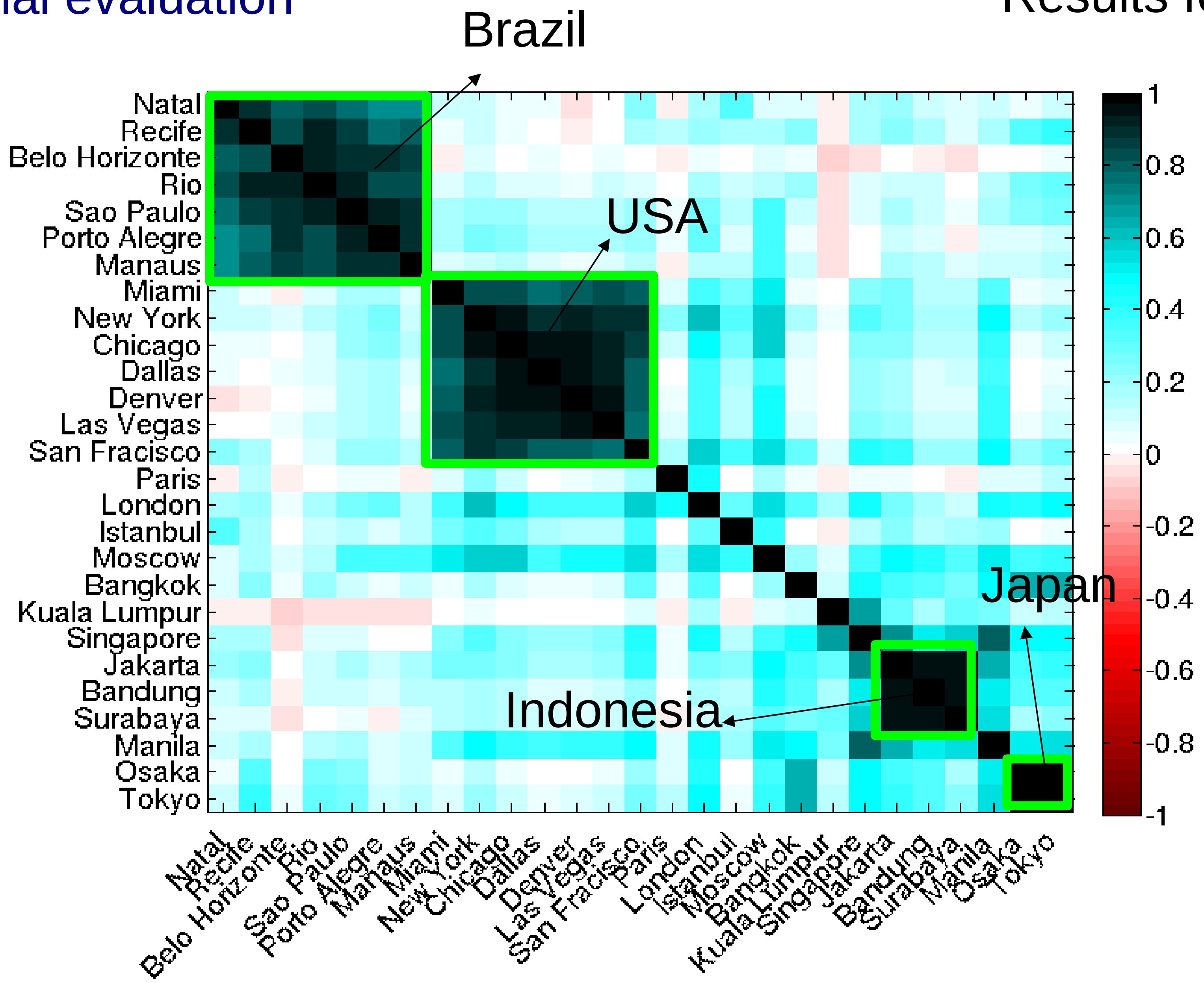
Results for **cities**



Extraction of cultural signatures

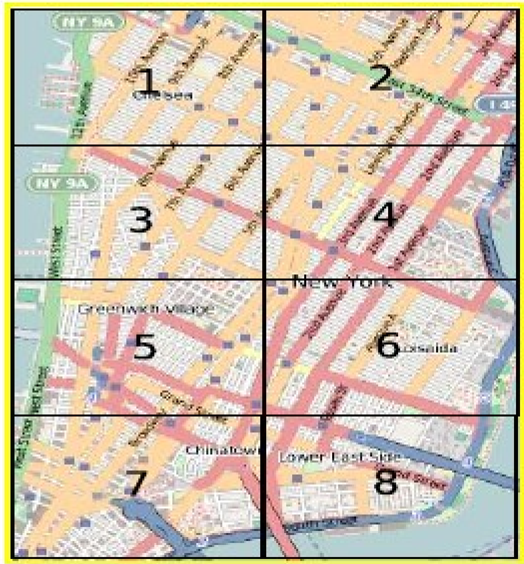
Spatial evaluation

Results for **cities**

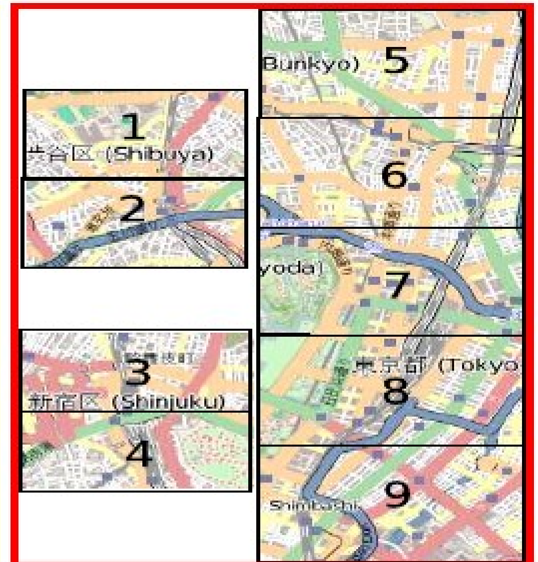


Extraction of cultural signatures

Spatial evaluation



New York (NY)



Tokyo (TKO)

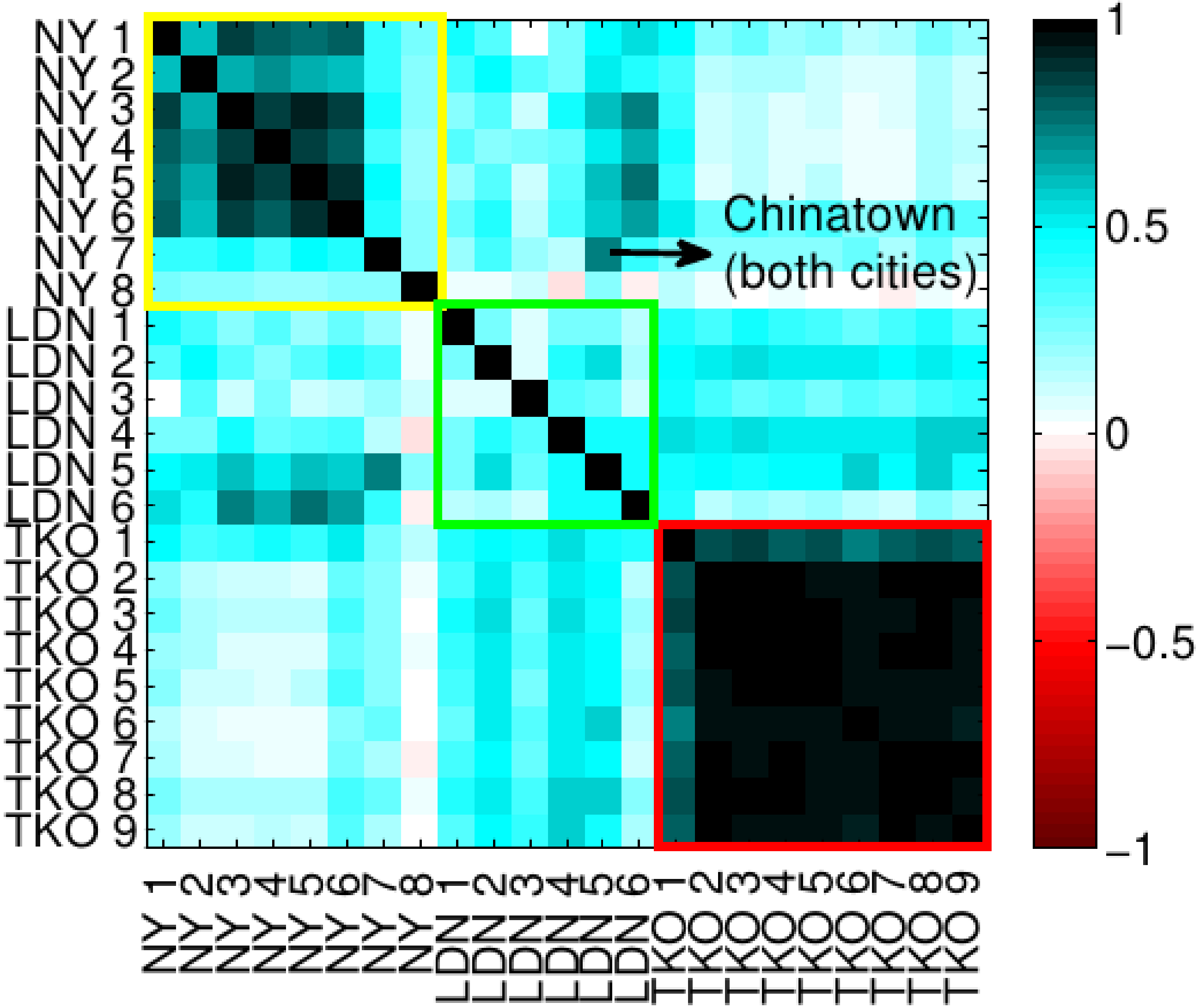
Popular areas

London (LND)

Extraction of cultural signatures

Spatial evaluation

Results for **areas inside cities**

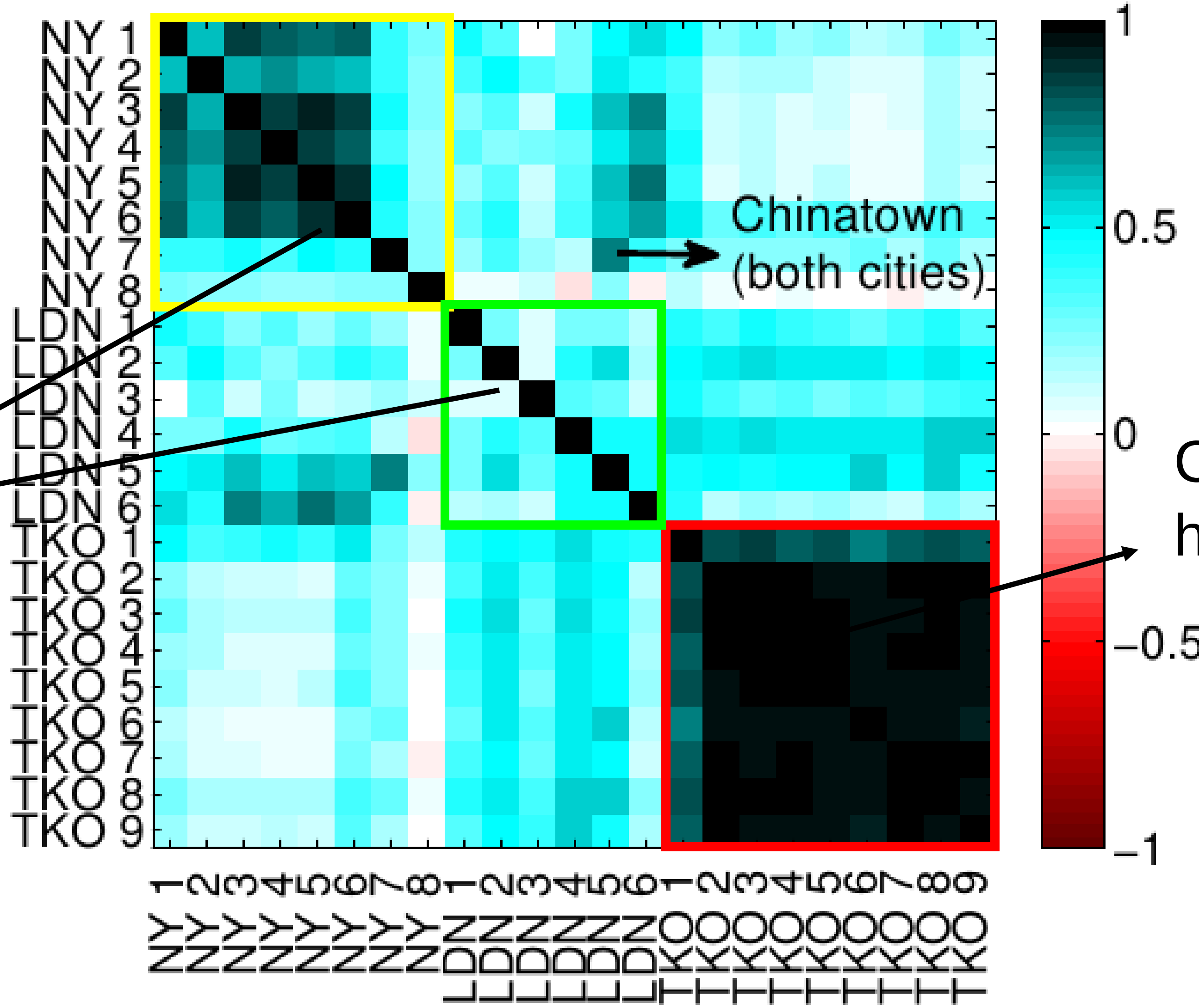


Extraction of cultural signatures

Spatial evaluation

Results for **areas inside cities**

Very international

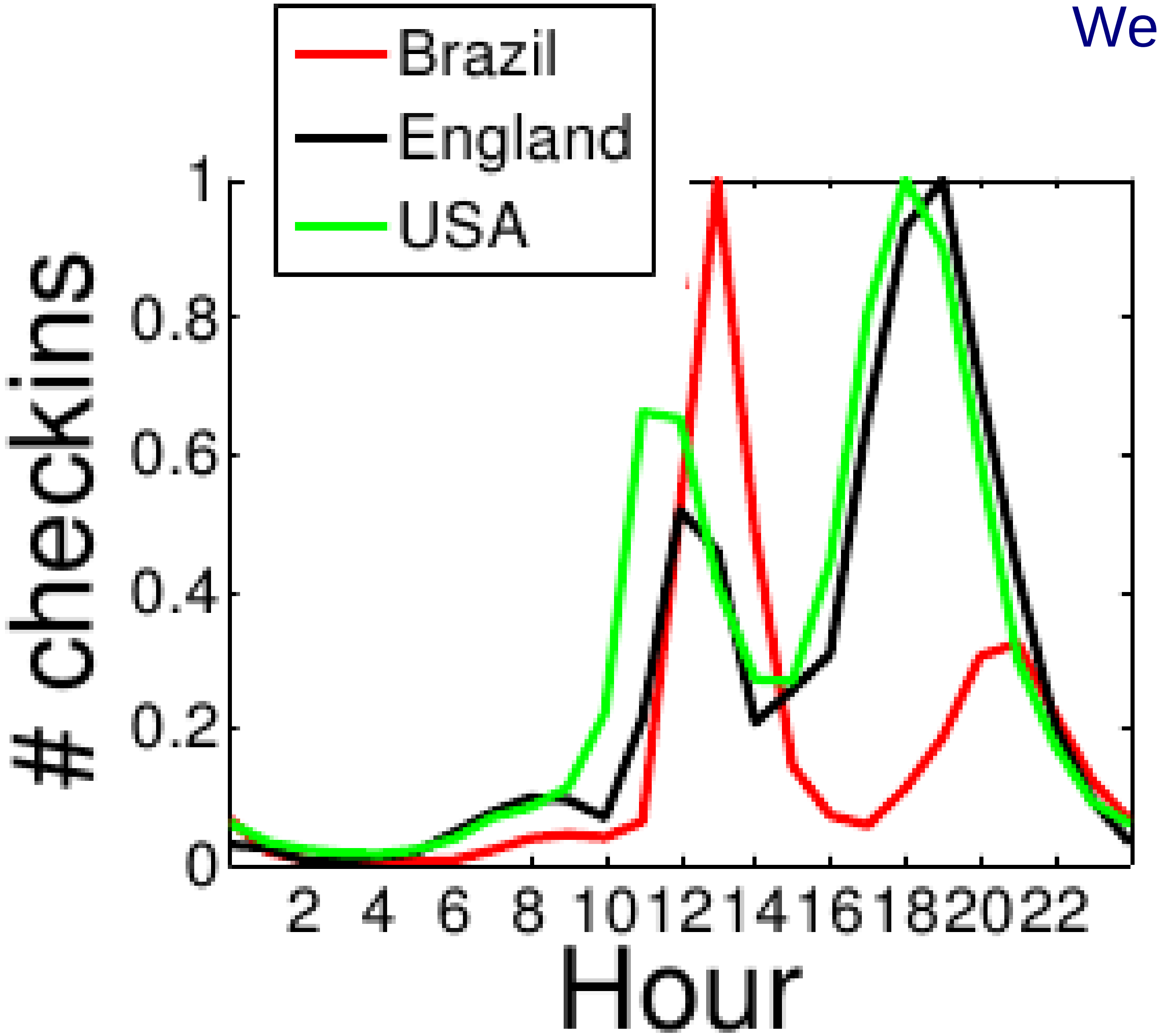


Culturally homogeneous

Extraction of cultural signatures

Temporal evaluation

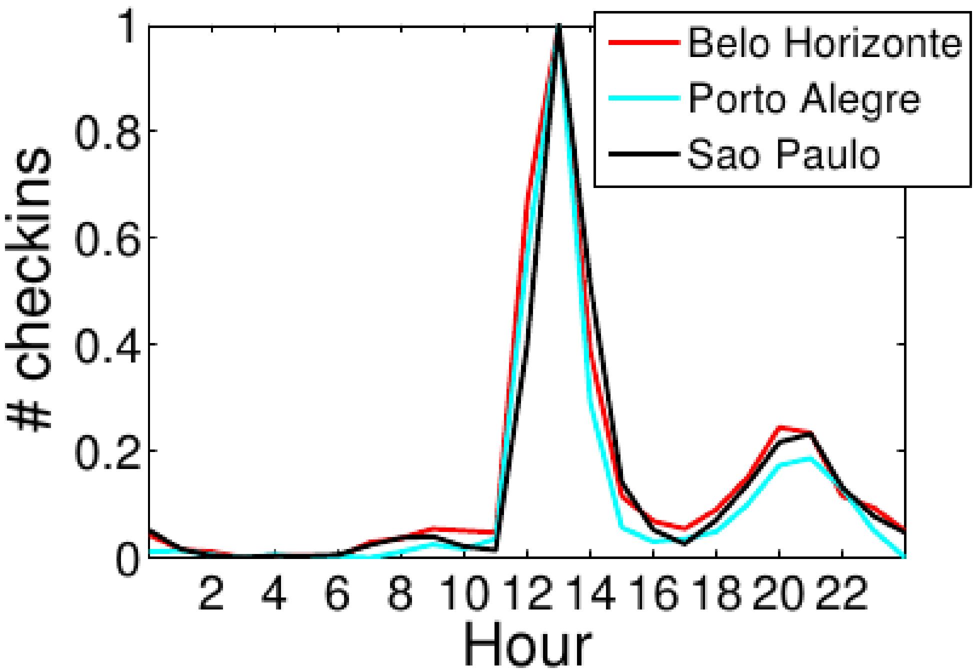
Weekdays



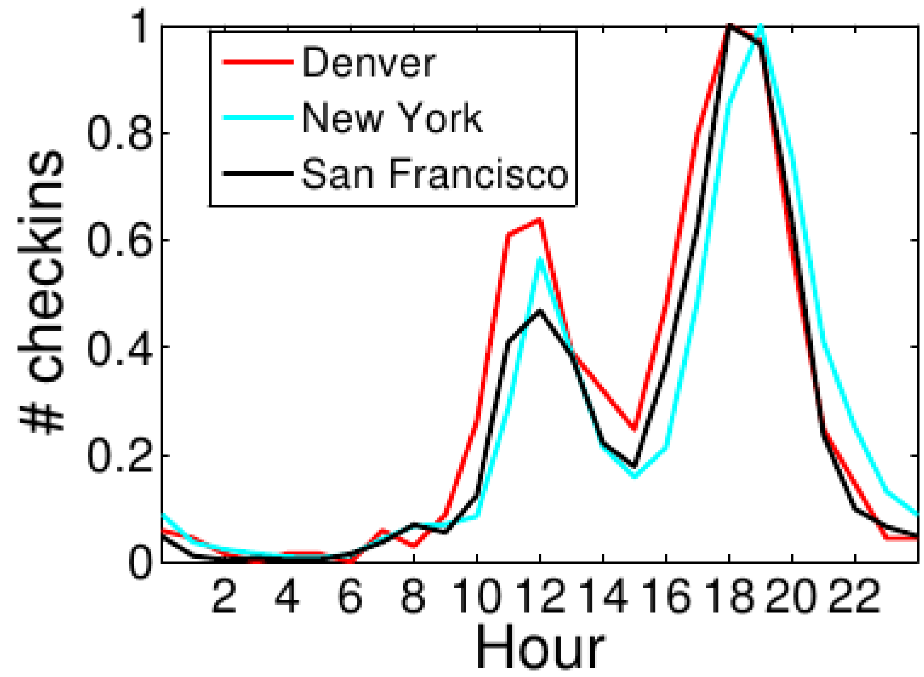
Extraction of cultural signatures

Temporal evaluation

Weekdays



Brazilian cities



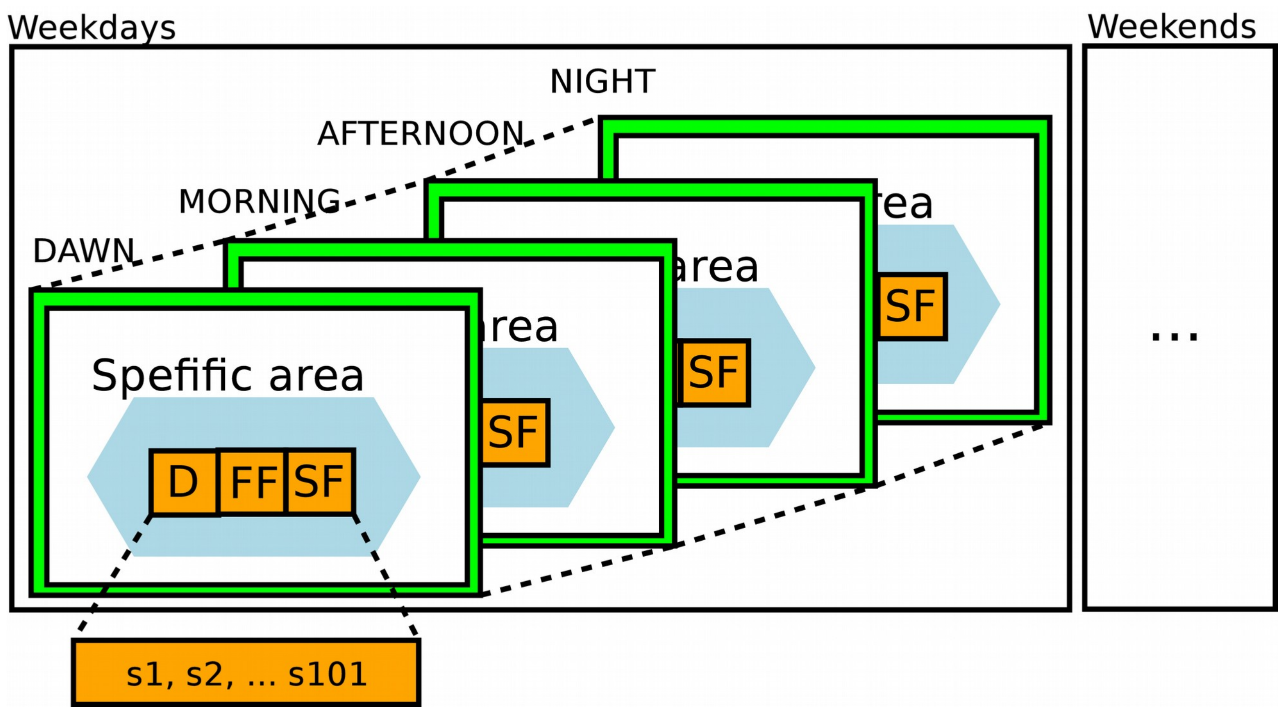
American cities

Most of the cities follow the general pattern of the country

Extraction of cultural signatures

Considered features: spatial / temporal

- Each area ***a*** has a normalized preference vector in **4 disjoint periods of the day** and on **weekdays** and **weekends**



General preference vector

D=drink / **FF**=fast food / **SF**=slow food

Identifying cultural boundaries

Preference vector for area
(time and space)

Identifying cultural boundaries

Preference vector for area
(time and space)

↓

Principal Component Analysis (PCA)

Identifying cultural boundaries

Preference vector for area
(time and space)

↓

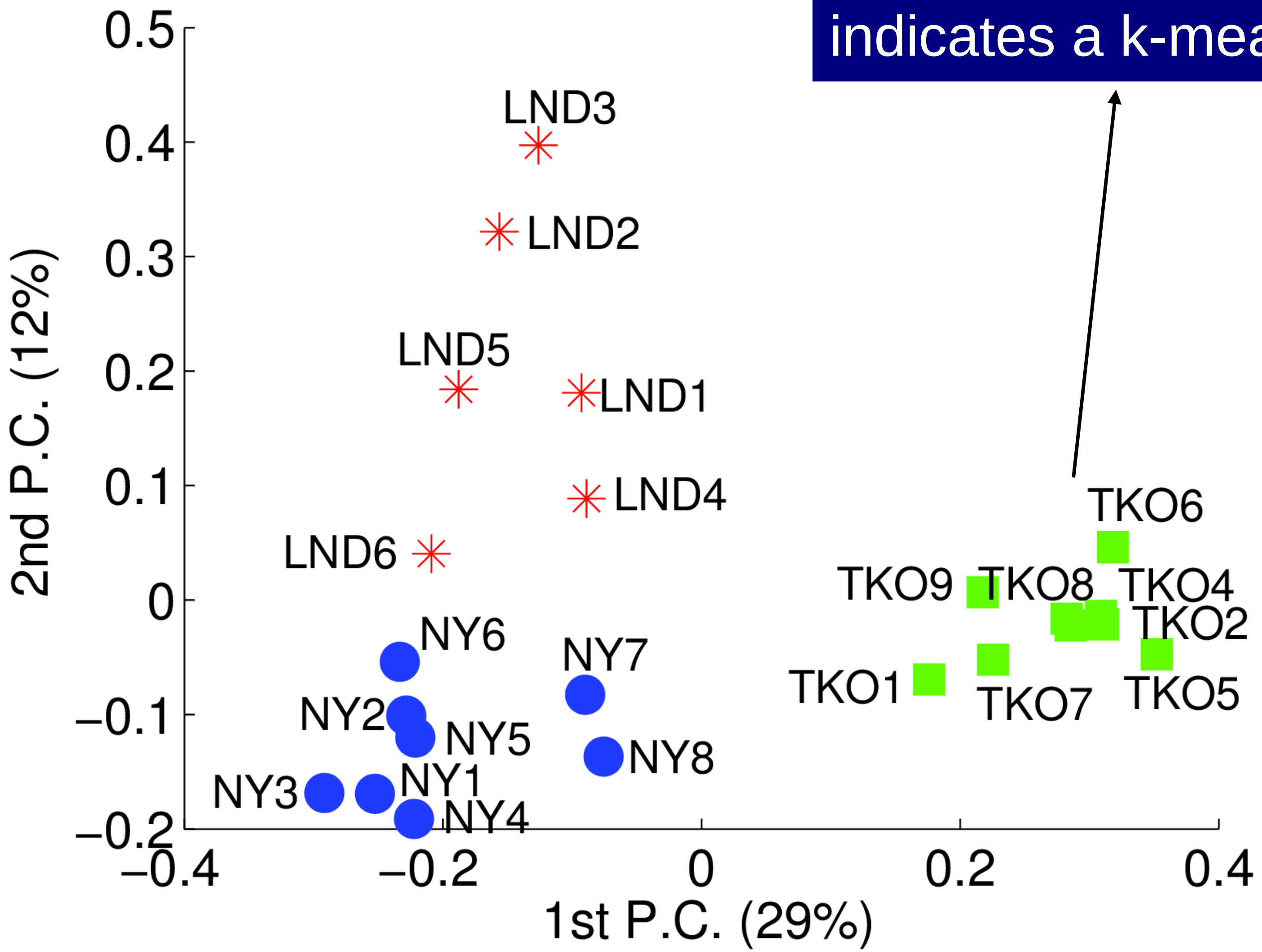
Principal Component Analysis (PCA)

k-means to group areas in the space defined by the PCs

Identifying cultural boundaries

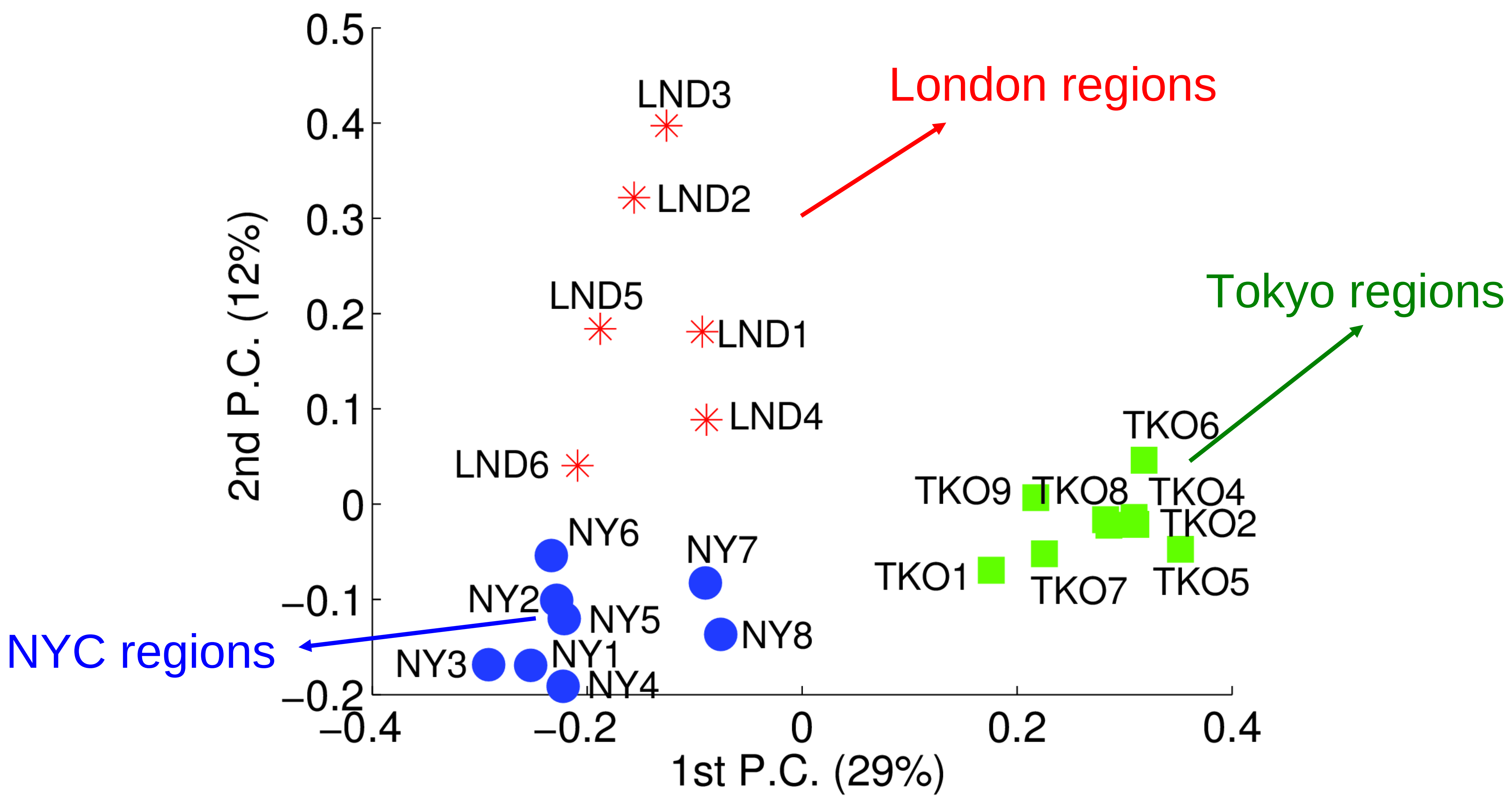
Clustering areas inside cities

Each color/symbol indicates a k-means cluster



Identifying cultural boundaries

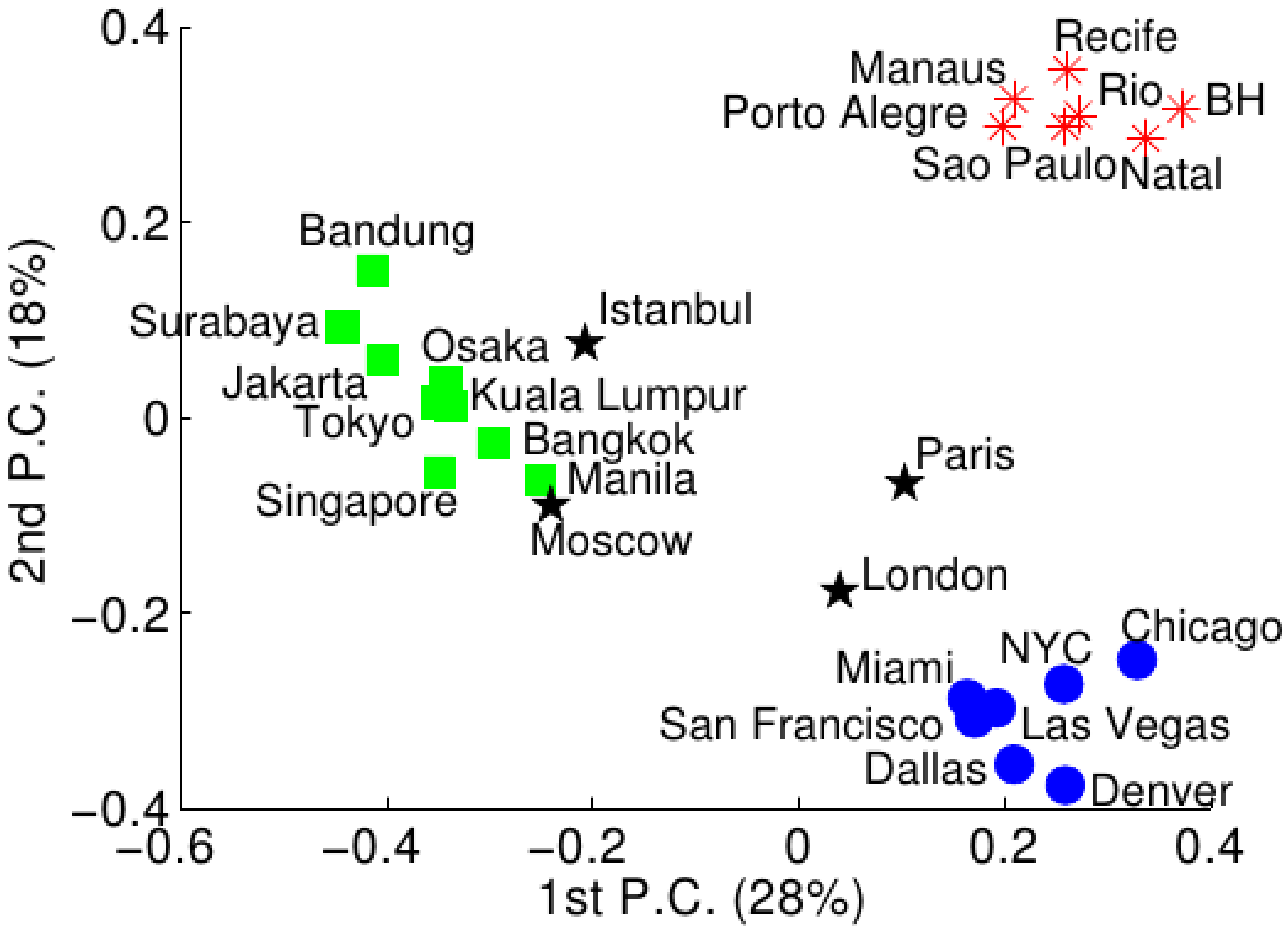
Clustering areas inside cities



k = 3 (3 different cities)

Identifying cultural boundaries

Clustering cities



k = 4 (4 distinct regions)

Identifying cultural boundaries

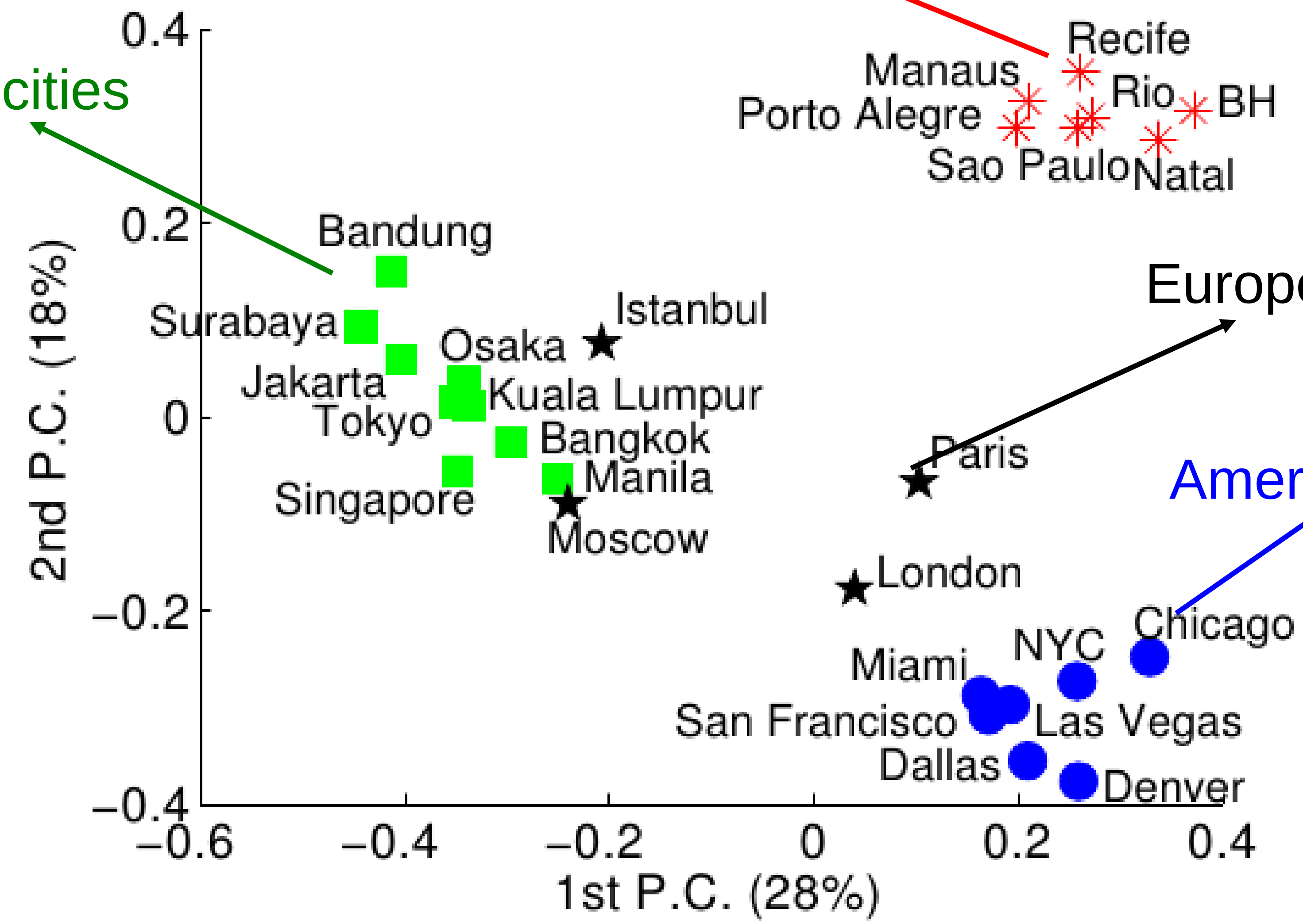
Clustering cities

Brazilian cities

Asian cities

European capitals

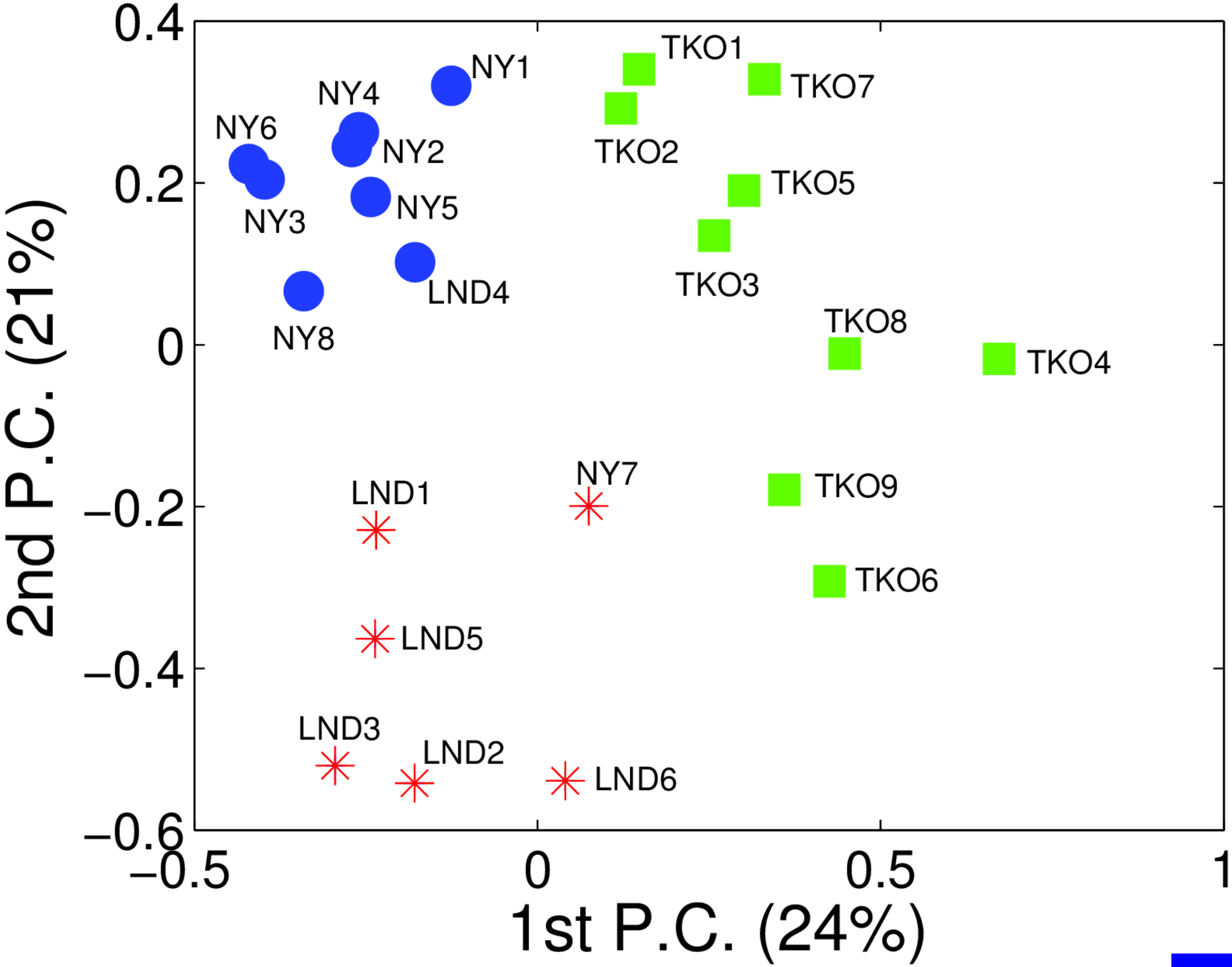
American cities



PSN Aplicability: cultural diferences

We can use a **partial set of the features** and **specific time**:

- E.G. Drink at weekend

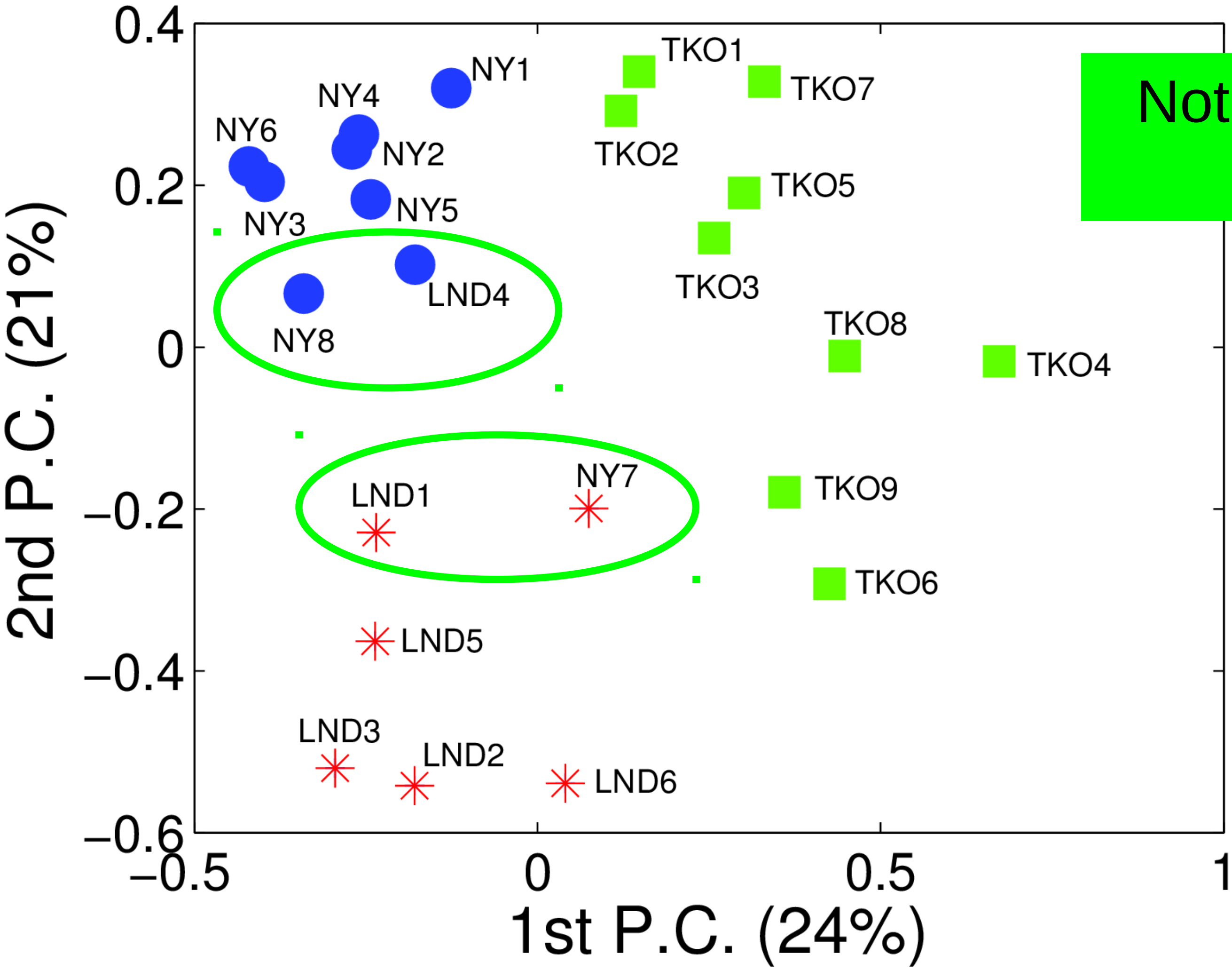


K = 3 (cities)

PSN Aplicability: cultural diferences

We can use a **partial set of the features** and **specific time**:

- E.G. Drink at weekend



Note the potential for area recommendation!

E outros traços culturais?

E outros traços culturais?

Features

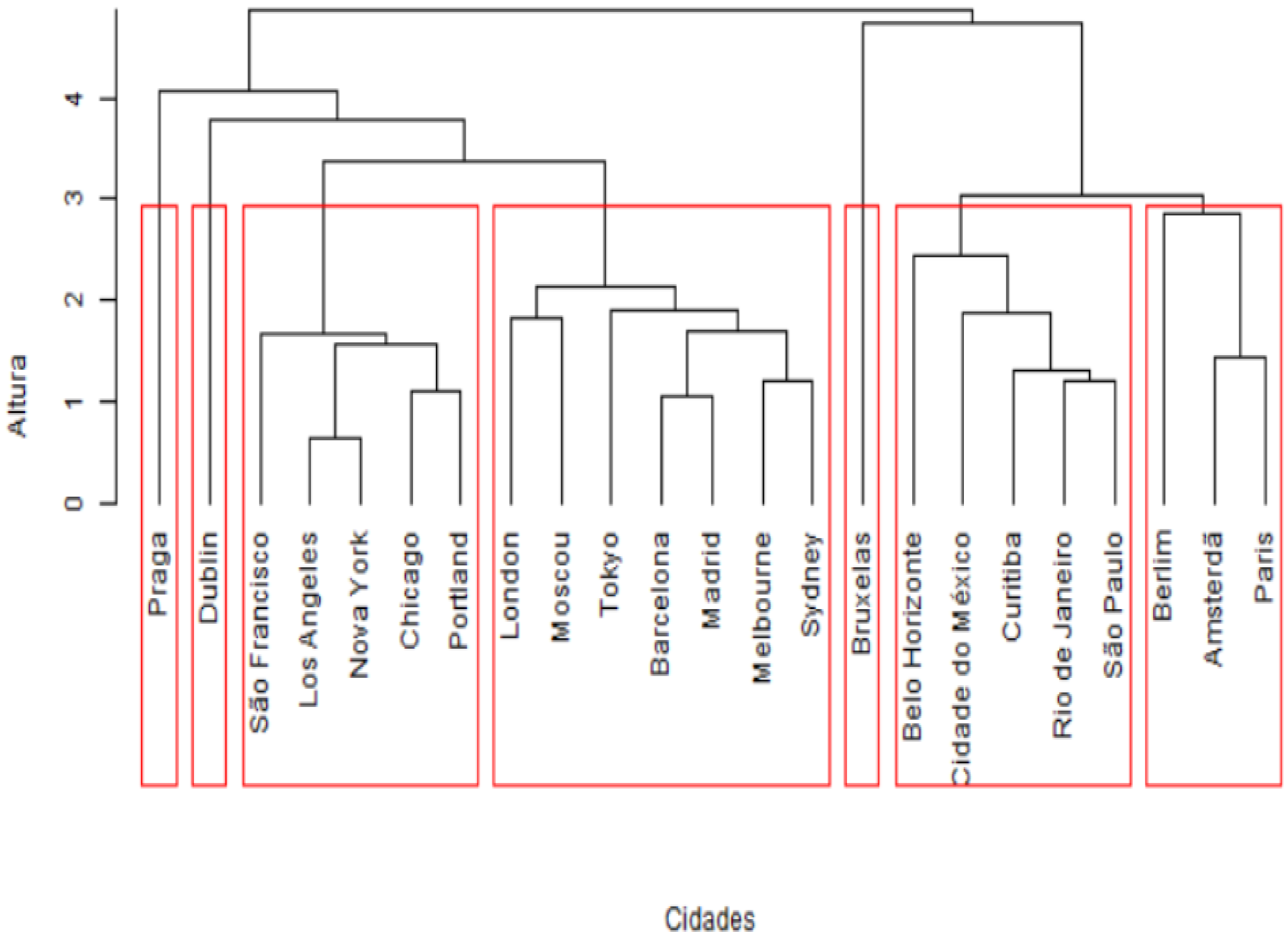
➔ Cada cidade c é representada por um vetor de preferências composto por classes de cervejas (features)

Cidade A

Cidade B

Resultados

➔ **Dendograma para o agrupamento realizado com as cidades**



Engagement of Polarized Groups


Jordan Kobellarz　　Alexandre Graeml

UTFPR

Michelle Reddy

Stanford University

- **Parrot Talk: Retweeting Among Twitter Users During the 2018 Brazilian Presidential Election**

Webmedia 2019

Focus

User engagement on Twitter regarding 2018 Brazil presidential election

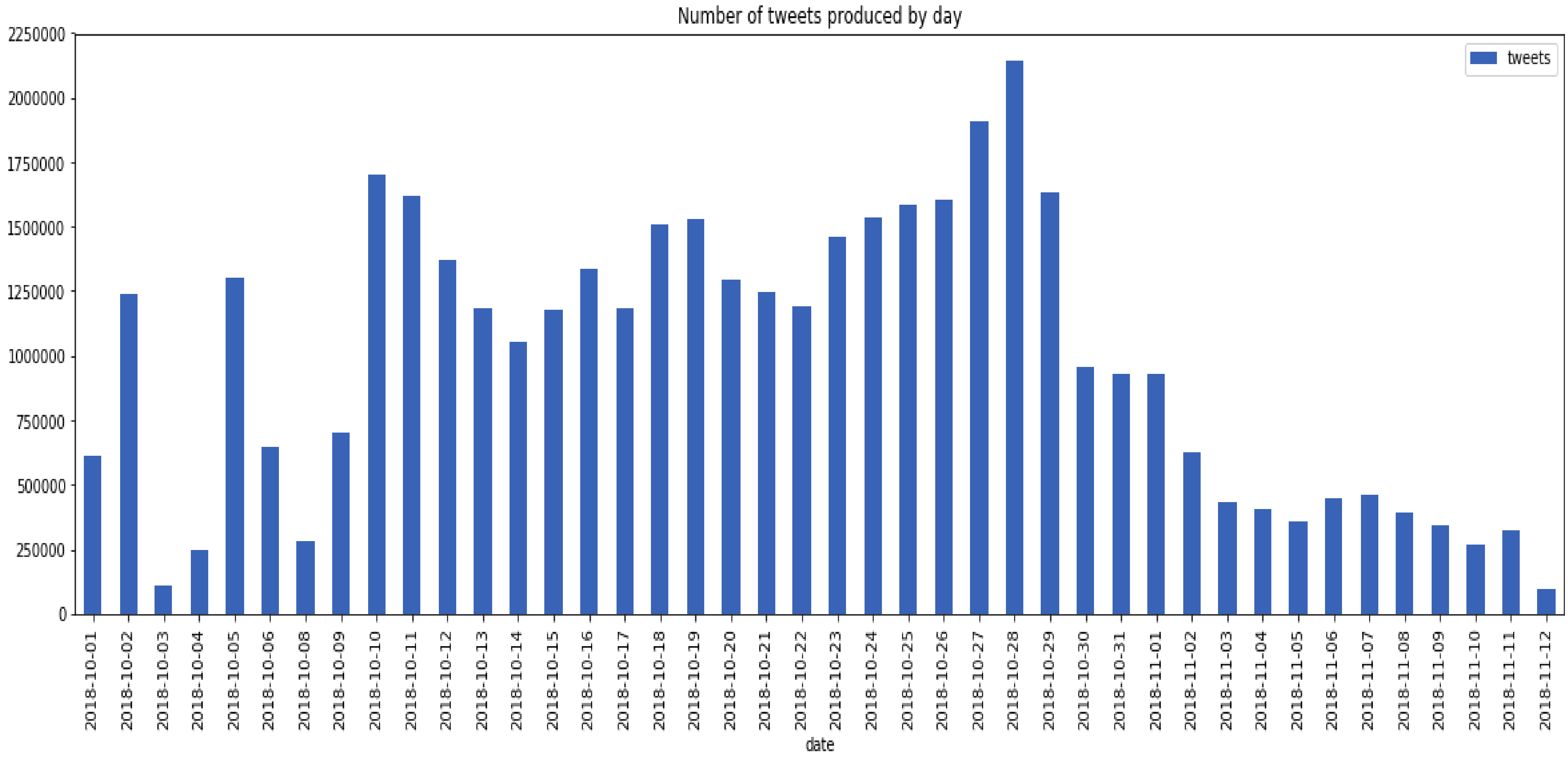
Data

Collection of

~ 36 Milion tweets

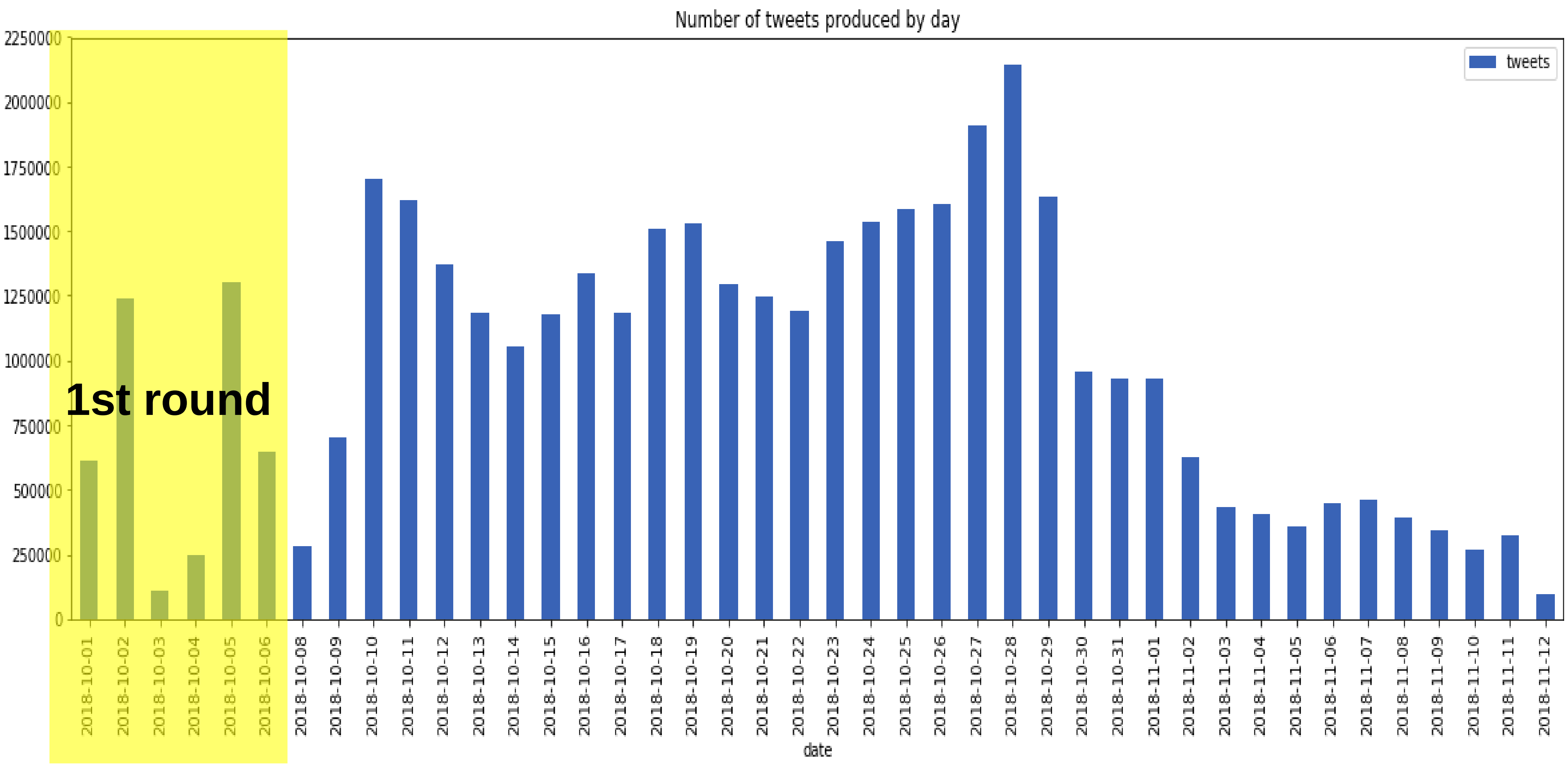
related to 2018 Brazilian Presidential Election

Data



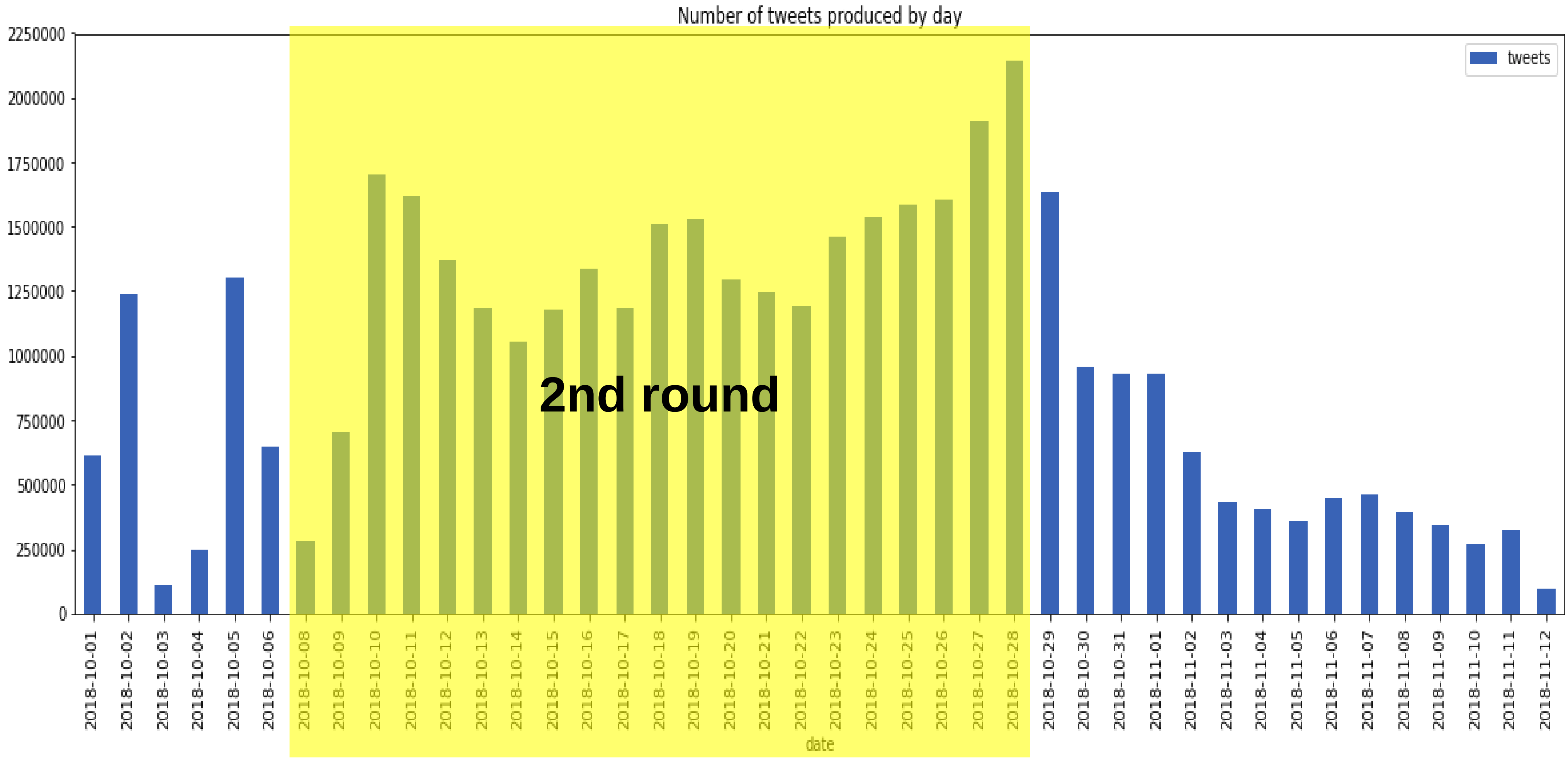
Number of tweets produced by day
tweets
0
250000
500000
750000
1000000
1250000
1500000
1750000
2000000
2250000
2018-10-01
2018-10-02
2018-10-03
2018-10-04
2018-10-05
2018-10-06
2018-10-08
2018-10-09
2018-10-10
2018-10-11
2018-10-12
2018-10-13
2018-10-14
2018-10-15
2018-10-16
2018-10-17
2018-10-18
2018-10-19
2018-10-20
2018-10-21
2018-10-22
2018-10-23
2018-10-24
2018-10-25
2018-10-26
2018-10-27
2018-10-28
2018-10-29
2018-10-30
2018-10-31
2018-11-01
2018-11-02
2018-11-03
2018-11-04
2018-11-05
2018-11-06
2018-11-07
2018-11-08
2018-11-09
2018-11-10
2018-11-11
2018-11-12
date

Data



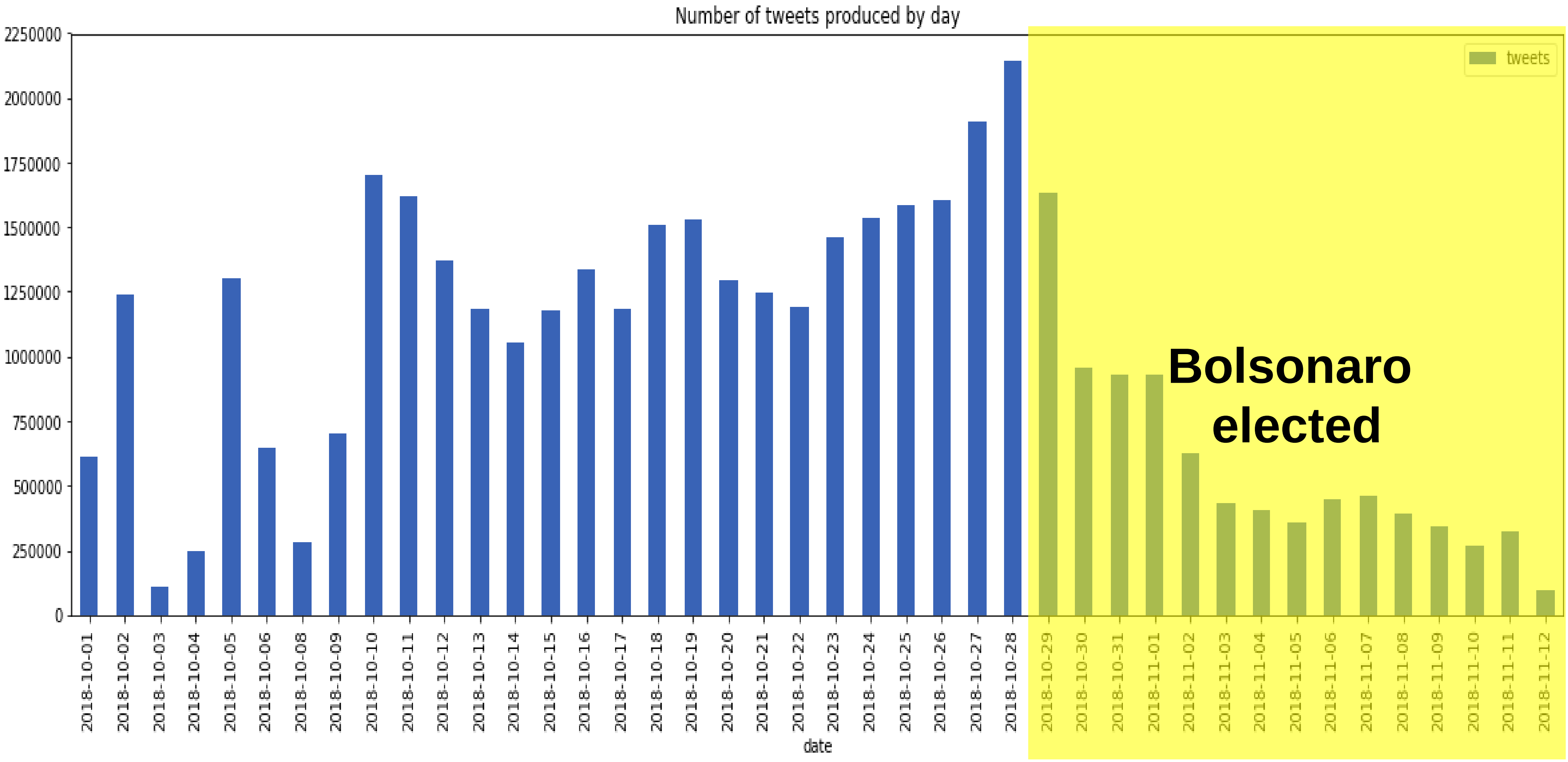
Number of tweets produced by day
tweets
1st round
date
0
250000
500000
750000
1000000
1250000
1500000
1750000
2000000
2250000
2018-10-01
2018-10-02
2018-10-03
2018-10-04
2018-10-05
2018-10-06
2018-10-08
2018-10-09
2018-10-10
2018-10-11
2018-10-12
2018-10-13
2018-10-14
2018-10-15
2018-10-16
2018-10-17
2018-10-18
2018-10-19
2018-10-20
2018-10-21
2018-10-22
2018-10-23
2018-10-24
2018-10-25
2018-10-26
2018-10-27
2018-10-28
2018-10-29
2018-10-30
2018-10-31
2018-11-01
2018-11-02
2018-11-03
2018-11-04
2018-11-05
2018-11-06
2018-11-07
2018-11-08
2018-11-09
2018-11-10
2018-11-11
2018-11-12

Data



Number of tweets produced by day
tweets
2nd round
date
0
250000
500000
750000
1000000
1250000
1500000
1750000
2000000
2250000
2018-10-01
2018-10-02
2018-10-03
2018-10-04
2018-10-05
2018-10-06
2018-10-08
2018-10-09
2018-10-10
2018-10-11
2018-10-12
2018-10-13
2018-10-14
2018-10-15
2018-10-16
2018-10-17
2018-10-18
2018-10-19
2018-10-20
2018-10-21
2018-10-22
2018-10-23
2018-10-24
2018-10-25
2018-10-26
2018-10-27
2018-10-28
2018-10-29
2018-10-30
2018-10-31
2018-11-01
2018-11-02
2018-11-03
2018-11-04
2018-11-05
2018-11-06
2018-11-07
2018-11-08
2018-11-09
2018-11-10
2018-11-11
2018-11-12

Data


Number of tweets produced by day
2250000
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
tweets
Bolsonaro elected
date
2018-10-01
2018-10-02
2018-10-03
2018-10-04
2018-10-05
2018-10-06
2018-10-08
2018-10-09
2018-10-10
2018-10-11
2018-10-12
2018-10-13
2018-10-14
2018-10-15
2018-10-16
2018-10-17
2018-10-18
2018-10-19
2018-10-20
2018-10-21
2018-10-22
2018-10-23
2018-10-24
2018-10-25
2018-10-26
2018-10-27
2018-10-28
2018-10-29
2018-10-30
2018-10-31
2018-11-01
2018-11-02
2018-11-03
2018-11-04
2018-11-05
2018-11-06
2018-11-07
2018-11-08
2018-11-09
2018-11-10
2018-11-11
2018-11-12

Classification of Hashtags

114.512

Unique hashtags

Classification of Hashtags

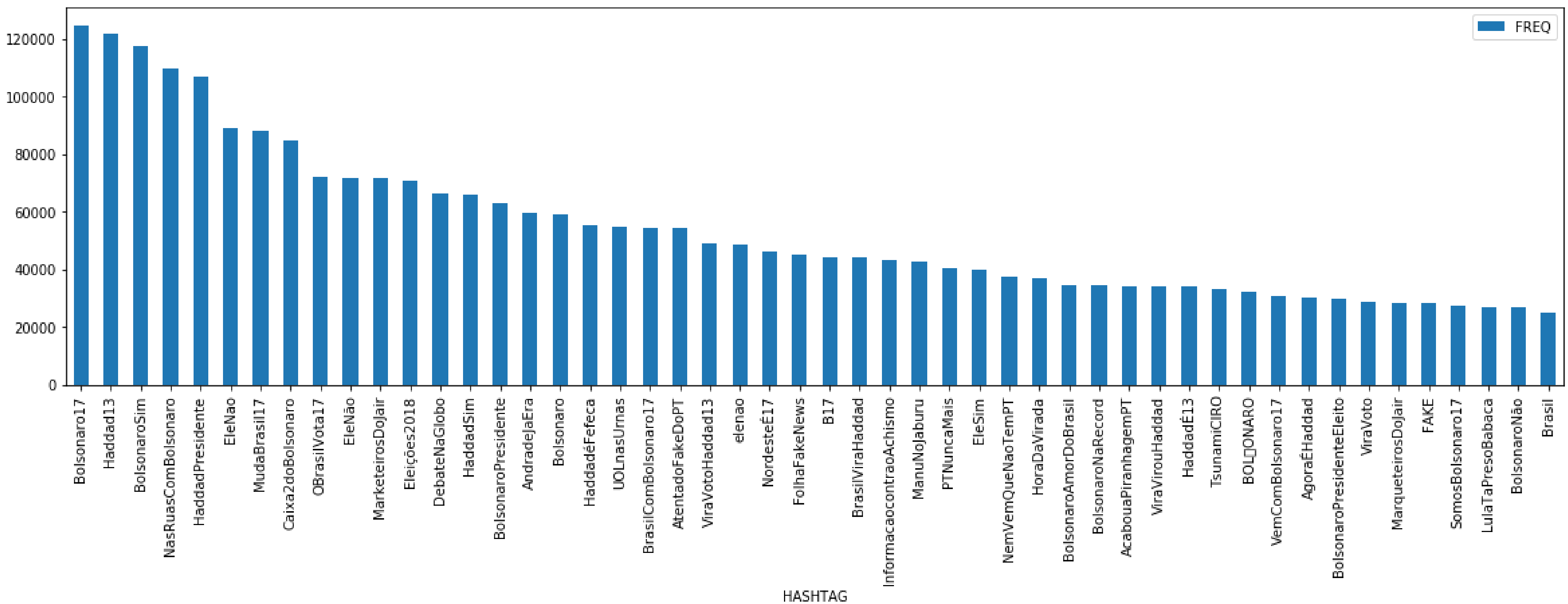
114.512

Unique hashtags

Hard do classify manually!

Classification of Hashtags

100 most popular



Classification of Hashtags

Help of volunteers

#AndradeJaEra, #Eleições2018, #FolhaFakeNews, #UOLnasUrnas, #Brasil, #InformacaocontraoAchismo, #HoraDaVirada, #VemProDebate, #ViraVoto, #VotoEmCedula, #RodaViva, #AFalhaéCafonérrima, #G1, #SeuVotoMePõeEmRisco, #FAKE, #NãoAceitaremosFraude, #SanatórioGeral, #EAgoraTSE, #Folha, #NaoAceitaremosFraude, #IbopeFake, #viravoto, #democracia, #SuasticaFake, #FakeNews, #Brazil, #SomosTodosReginaDuarte, #delegadofrancischini, #Eleicoes2018 e #BrasilDecide.

#NasRuasComBolsonaro, #BolsonaroSim, #MudaBrasil17, #Bolsonaro17, #Bolsonaro, #BrasilComBolsonaro17, #NordesteÉ17, #BolsonaroPresidente, #NemVemQueNaoTemPT, #PTNuncaMais, #BolsonaroPresidenteEleito, #LulaTaPresoBabaca, #HaddadNãoÉCristão, #B17, #Nordeste17, #OLulaTaPresoBabaca, #EleSim, #bolsonaro17, #PTNão, #PTnão, #FolhaP******DoPT (Conteúdo impróprio), #AcabouaPiranhagemPT, #BolsonaroPresidente17, #elesim, #bolsonaro e #PTnao.

#Haddad13, #Caixa2doBolsonaro, #HaddadPresidente, #HaddadSim, #EleNao, #BrasilViraHaddad, #EleNão, #AgoraÉHaddad, #BolsonaroNão, #ViraVotoHaddad13, #CassaçãoDoBolsonaro, #LulaLivre, #bolsonaroCagao, #Caixa2DoBolsonaro, #Haddad, #elenao, #ViraVirouHaddad13, #ELENAO, #MaisLivrosMenosArmas e #haddadpresidente.

Classification of Hashtags

76 agreements

Uncertain (?)

#AndradeJaEra, #Eleições2018, #FolhaFakeNews, #UOLnasUrnas, #Brasil, #InformacaocontraoAchismo, #HoraDaVirada, #VemProDebate, #ViraVoto, #VotoEmCedula, #RodaViva, #AFalhaéCafonérrima, #G1, #Se[illegible] #[illegible]AceitaremosFraude, #SanatórioGeral, #EAgoraTSE, #Folha, #NaoAceitaremosFraude, #IbopeFake, #viravoto, #democracia, #[illegible], #FakeNews, #Brazil, #SomosTodosReginaDuarte, #delegadofrancischini, #Eleicoes2018 e #BrasilDecide.

Right (R)

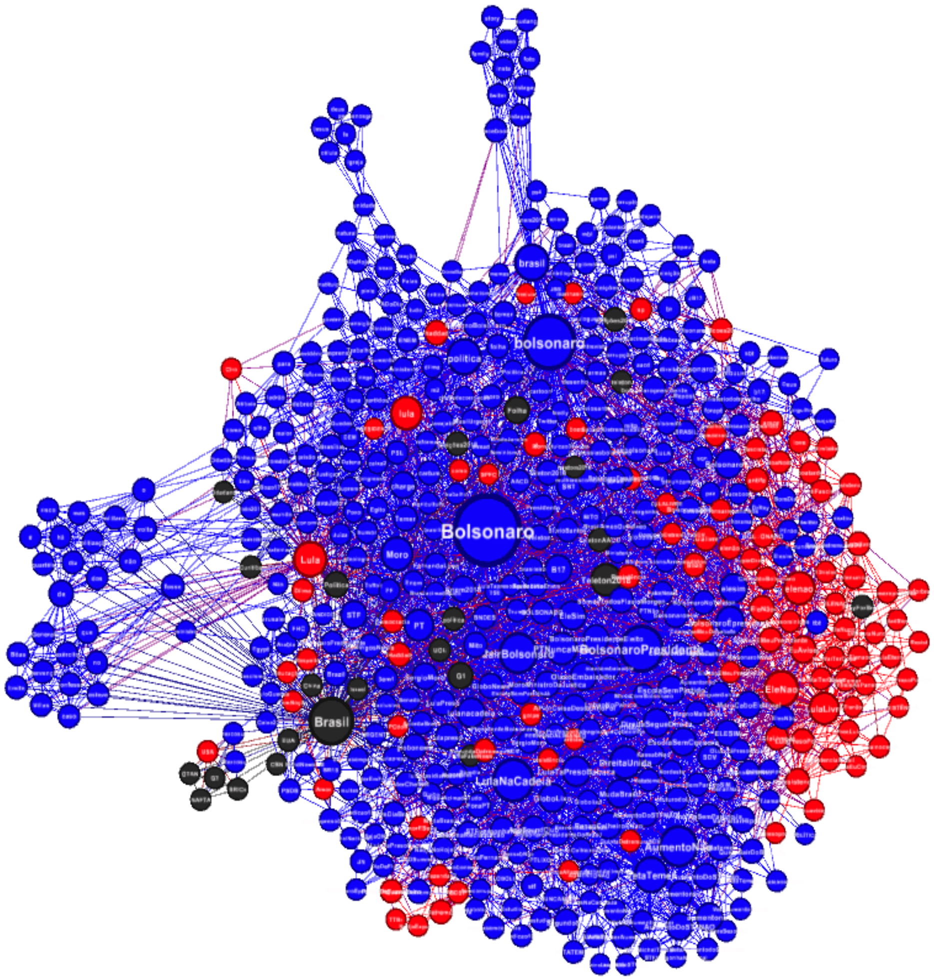
#NasRuasComBolsonaro, #BolsonaroSim, #MudaBrasil17, #Bolsonaro17, #Bolsonaro, #BrasilComBolsonaro17, #NordesteÉ17, #BolsonaroPresidente, #NemVemQueNaoTemPT, #PTNuncaMais, #BolsonaroPresidenteEleito, #LulaTaPresoBabaca, #HaddadNãoÉCristão, #B17, #Nordeste17, #OLulaTaPresoBabaca, #EleSim, #bolsonaro17, #PTNão, #PTnão, #FolhaP***** [illegible]prio), #AcabouaPiranhagemPT, #BolsonaroPresidente17, #elesim, #bolsonaro e #PTnao.

Left (R)

#Haddad13, #Caixa2doBolsonaro, #HaddadPresidente, #HaddadSim, #EleNao, #BrasilViraHaddad, #EleNão, #AgoraÉHaddad, #BolsonaroNão, #ViraVotoHaddad13, #CassaçãoDoBolsonaro, #LulaLivre, #bolsonaroC[illegible], #C[illegible]Bolsonaro, #Haddad, #elenao, #ViraVirouHaddad13, #ELENAO, #MaisLivrosMenosArmas e #haddadpresidente.

Classification of Hashtags

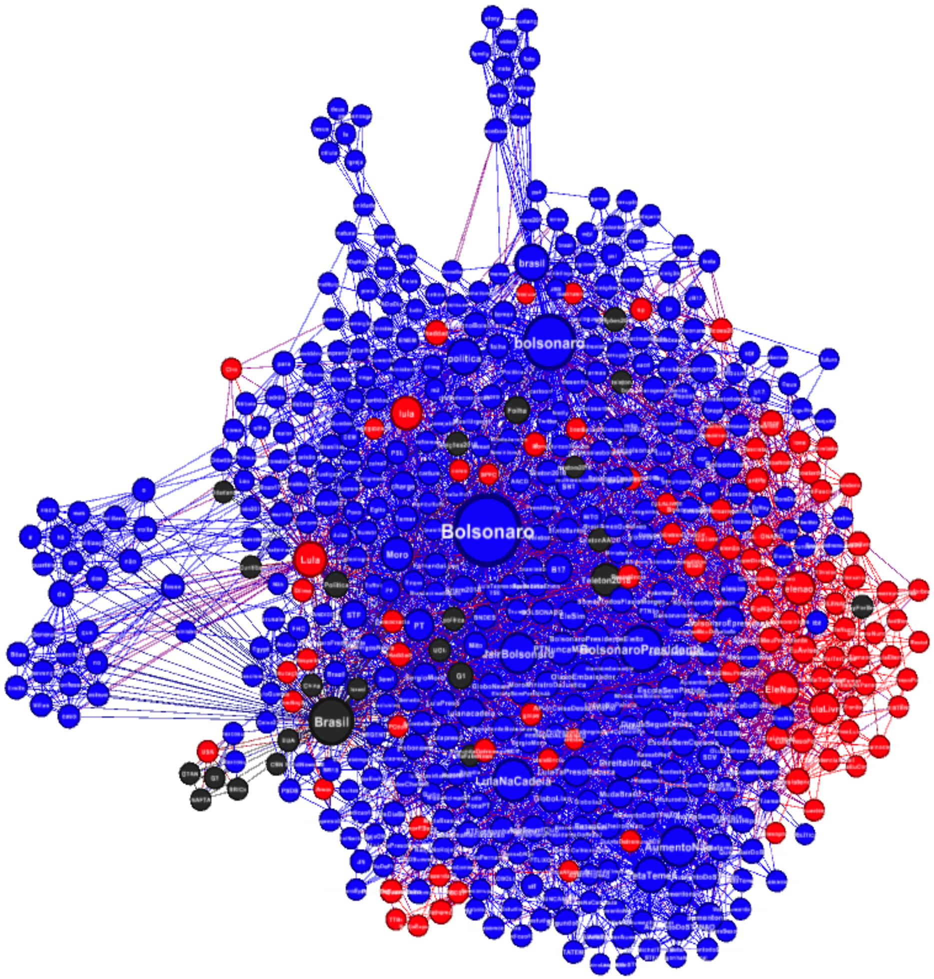
Network of co-occurrences of hashtags (all tweets)



Semi-supervised learning using gaussian fields and harmonic functions (Zhu, 2003)

Classification of Hashtags

Network of co-occurrences of hashtags (all tweets)



Semi-supervised learning using gaussian fields and harmonic functions (Zhu, 2003)

Classification of **78,649** hashtags (68.7% of total)

Only tweets that have one those

Classification of Users

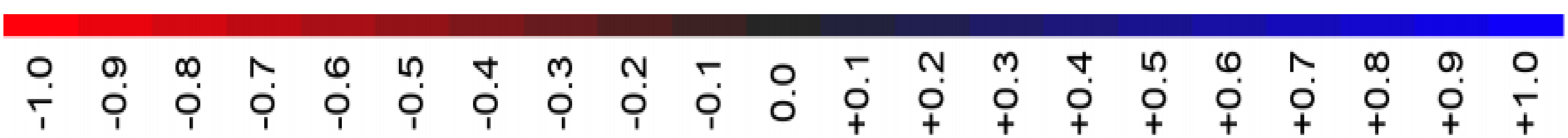
$$P(H) = \frac{|H_R| - |H_L|}{|H|},$$

$$H = H_R + H_L + H_?$$

Classification of Users

$$P(H) = \frac{|H_R| - |H_L|}{|H|},$$

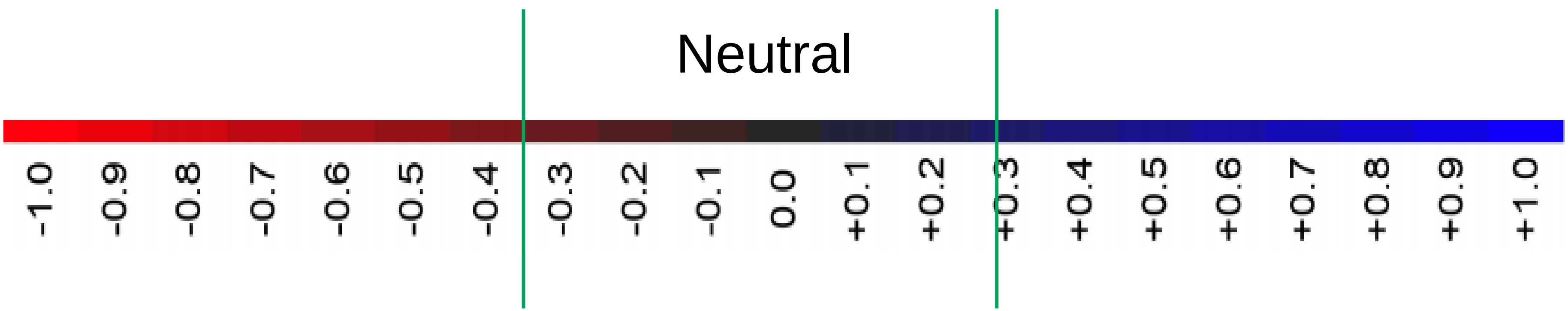
$$H = H_R + H_L + H_?$$



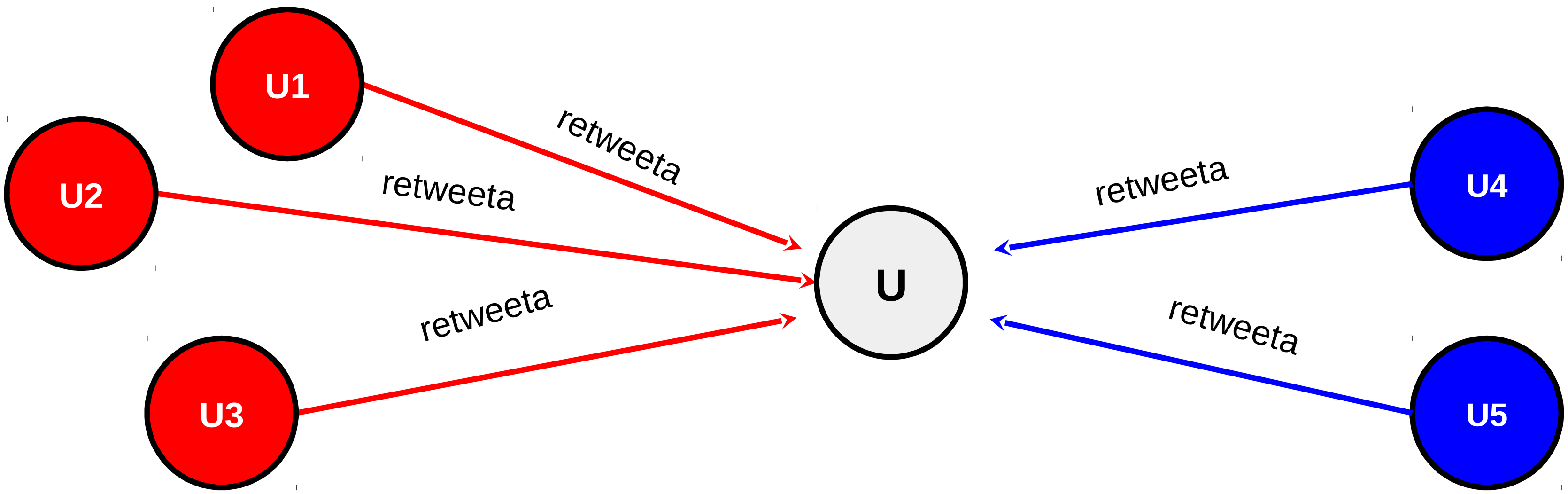
Classification of Users

$$P(H) = \frac{|H_R| - |H_L|}{|H|},$$

$$H = H_R + H_L + H_?$$

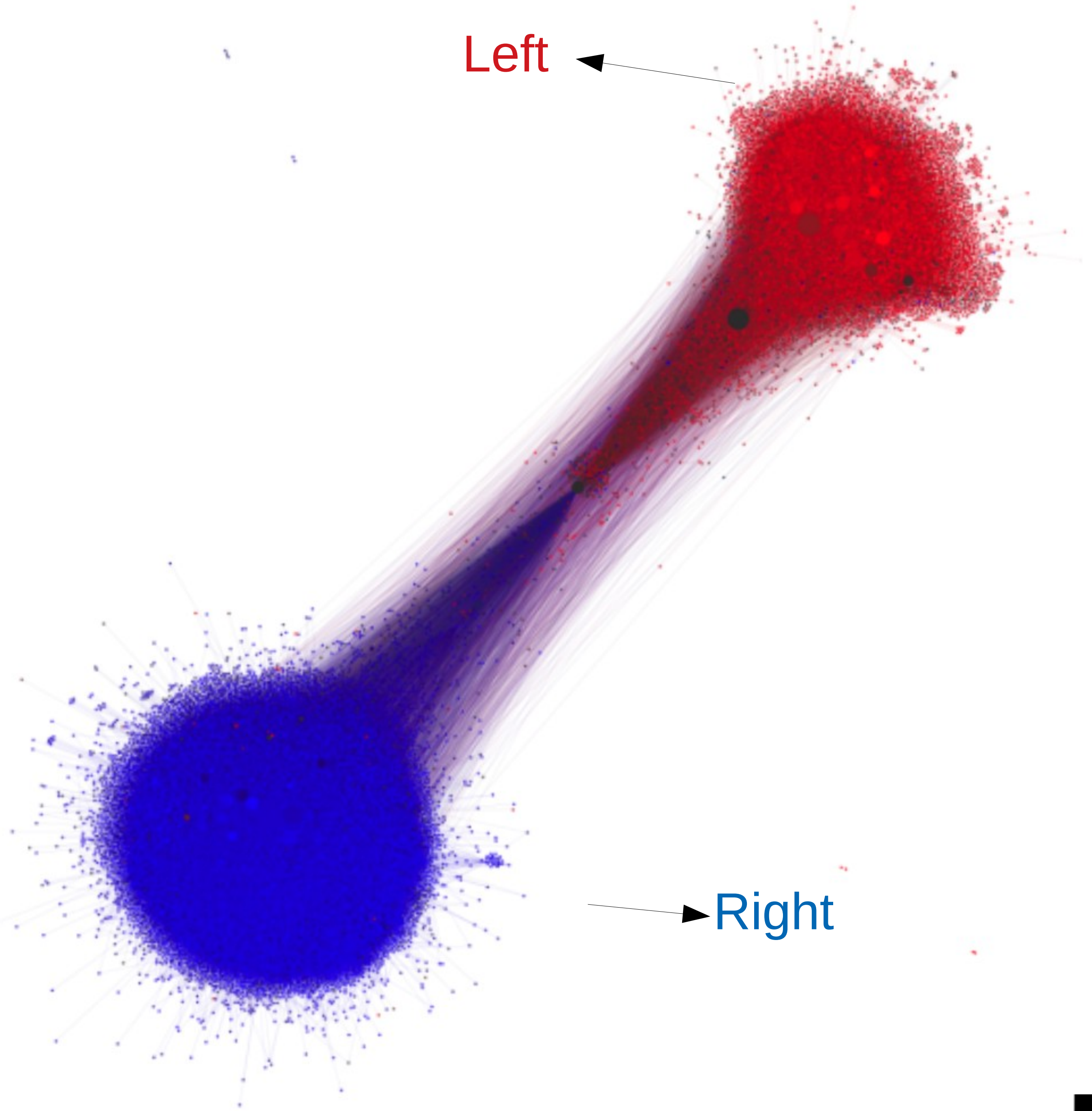


Engagement Graph



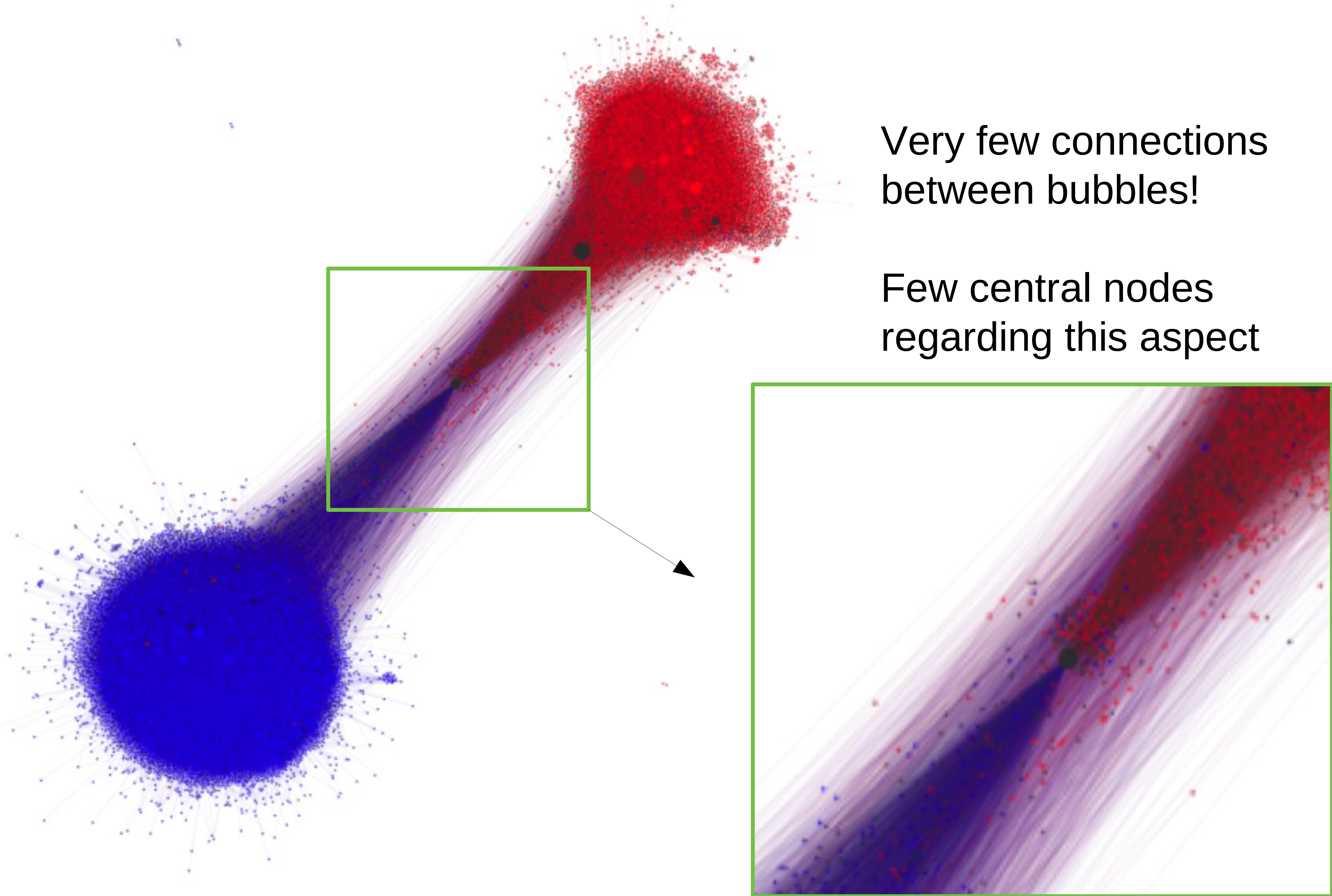
Weight represents the amount of retweets

Engagement Graph



"Echo Chambers"
[Barberá et al. 2015]

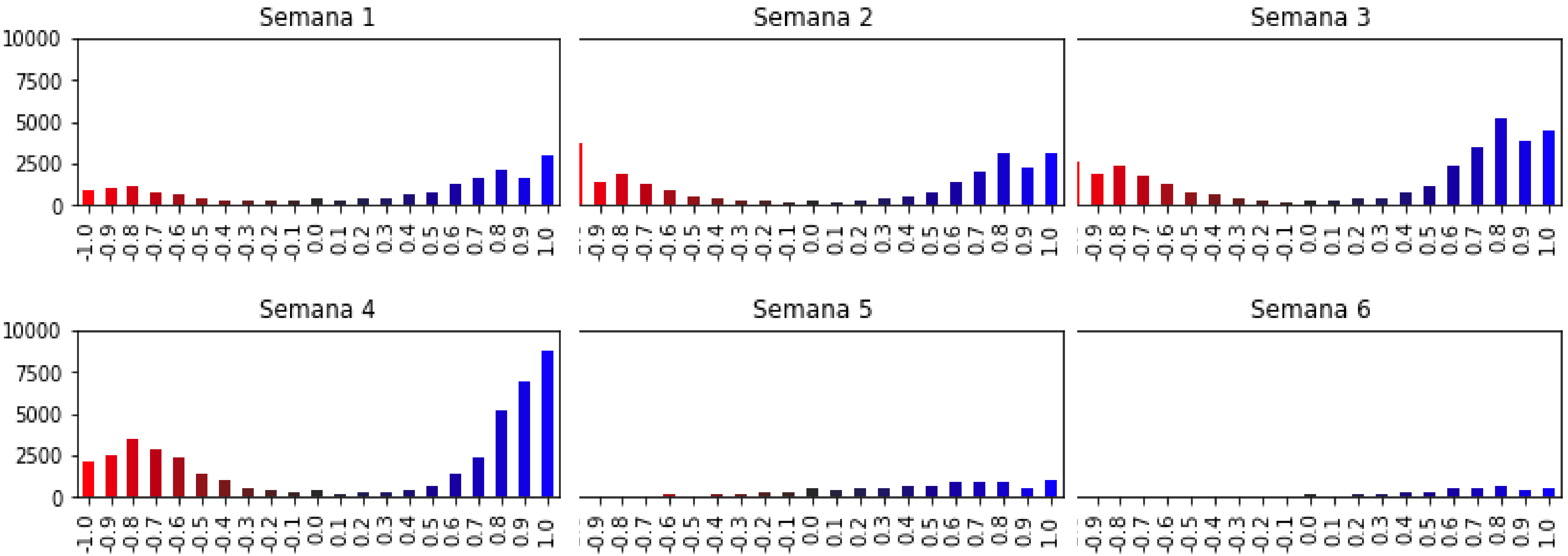
Engagement Graph



Very few connections between bubbles!

Few central nodes regarding this aspect

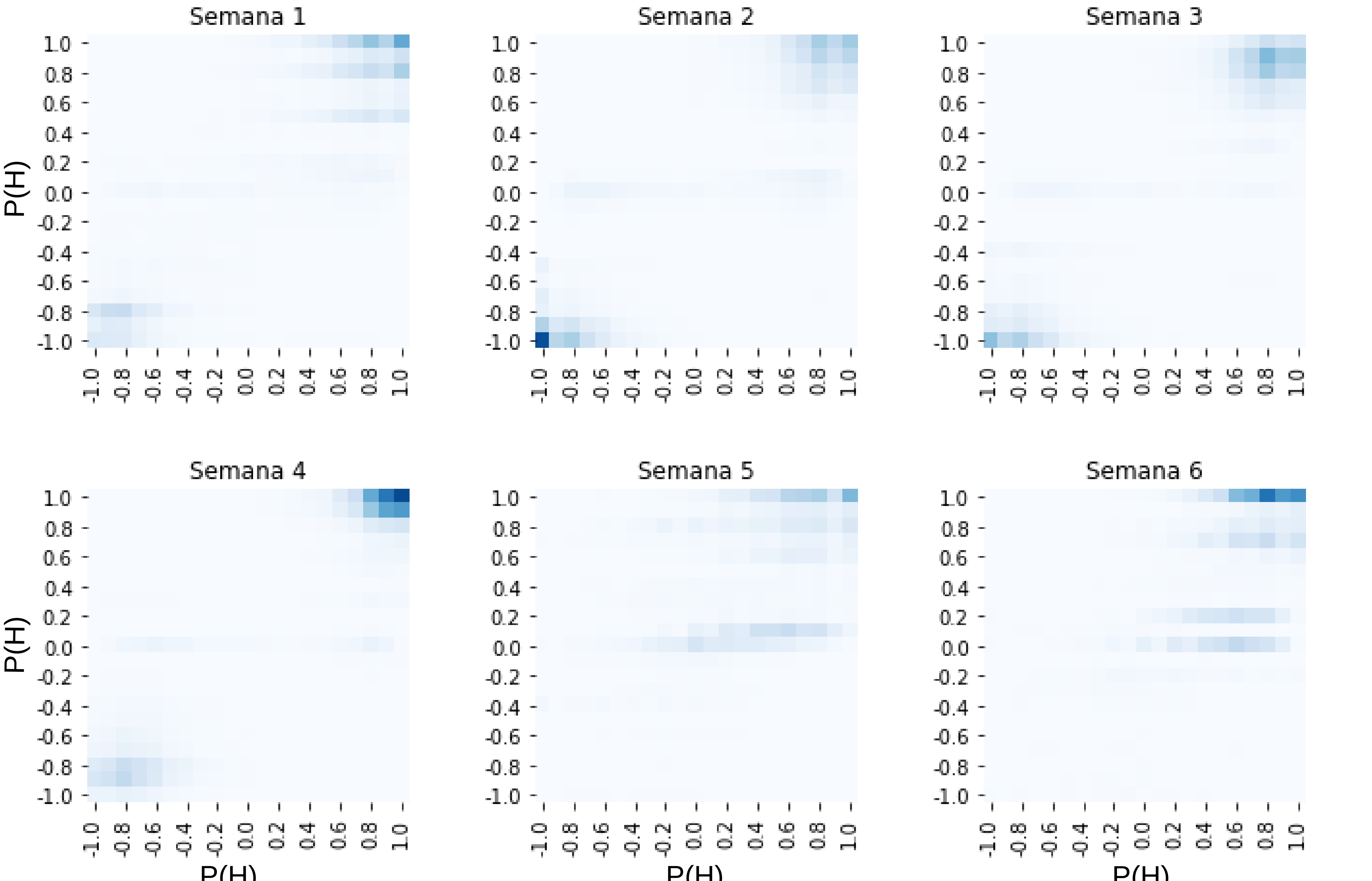
Polarity distribution



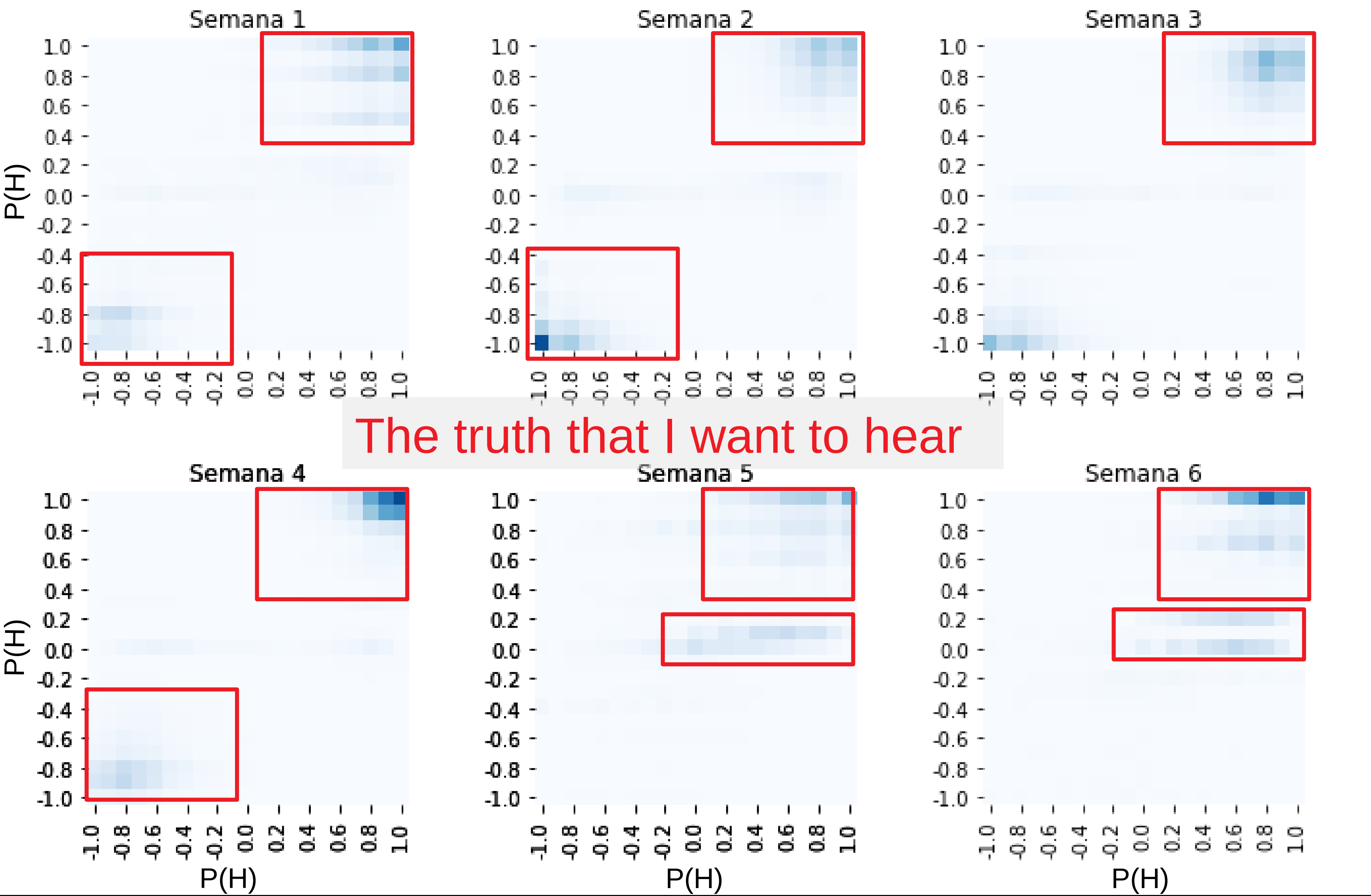
After the election

Binominal for polarization
[Fiorina e Abrams, 2008]

Interaction between users

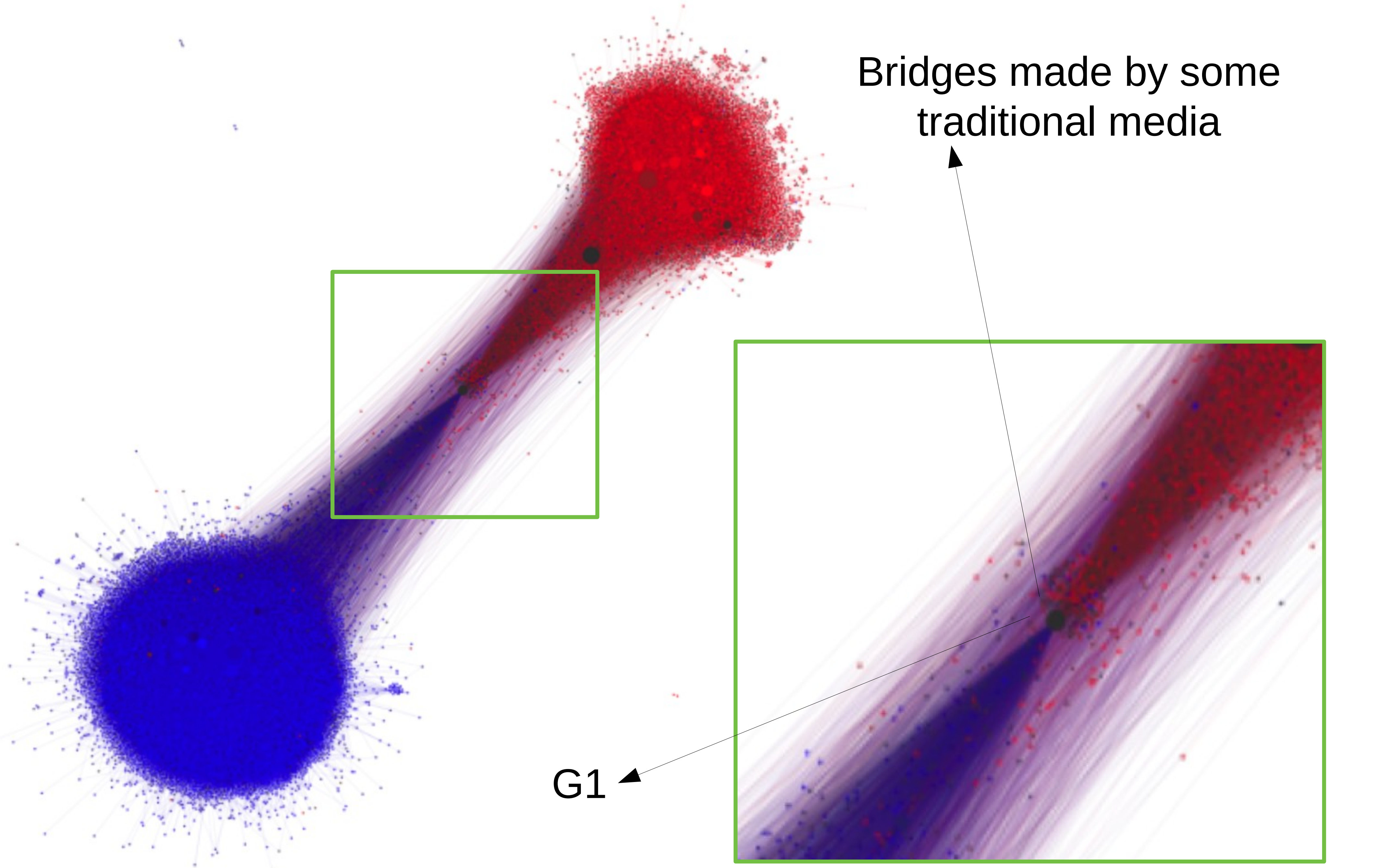


Interaction between users



The truth that I want to hear

Light in the end of the tunnel?



[redacted]

Ongoing research

Bridge mechanism

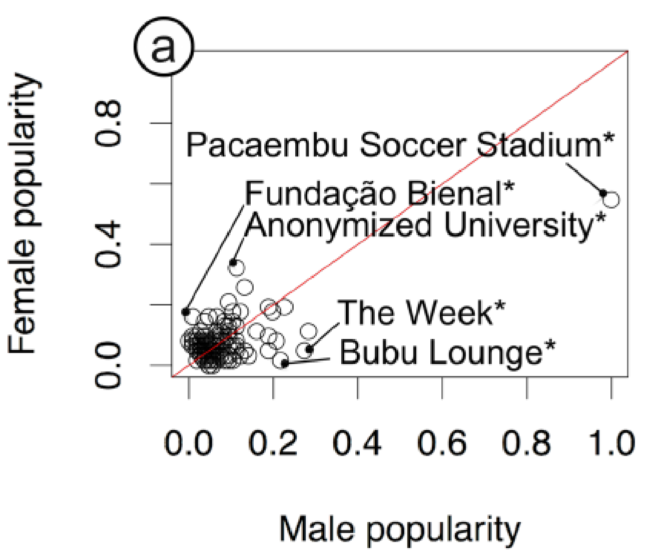
Several Phenomena Worth Investigating

Gender Differences

Willy Muller
(visiting student)

Method to quantify gender preferences in different regions

Identification of “anomalous” areas

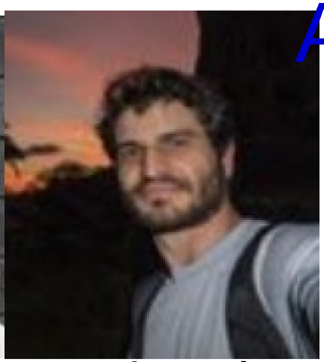


Gender Matters! Analyzing Global Cultural Gender Preferences for Venues Using Social Sensing
EPJ Data Science, 2017

Urban Planning and Place Branding

Giovani

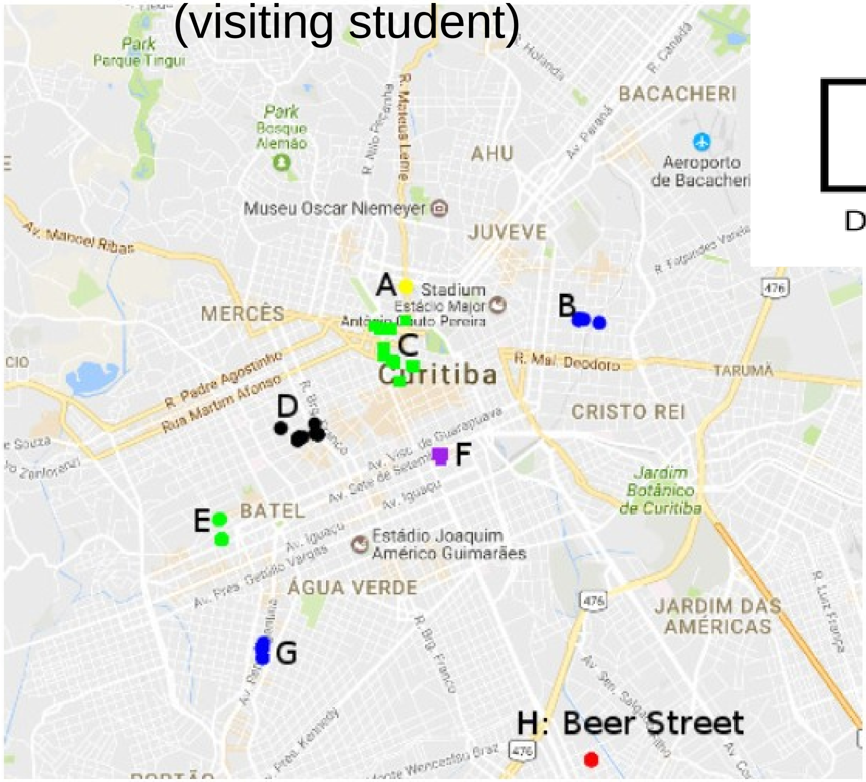
Ville Santalla (visiting student)

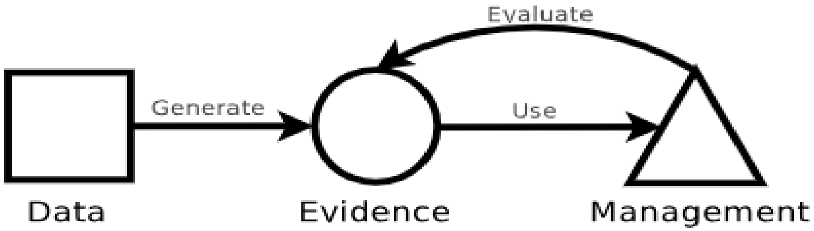


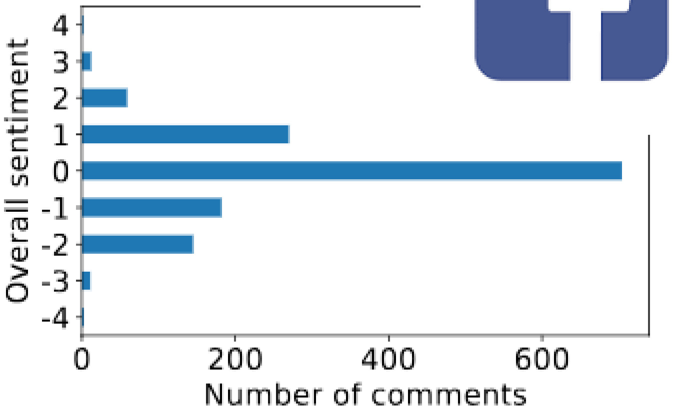
Luiz Celso

Adaptive and reactive planning (identifying hidden potential)

Things that are already happening!







Using Social Media to Improve the Management of Branding and Marketing of Cities: Findings from Curitiba

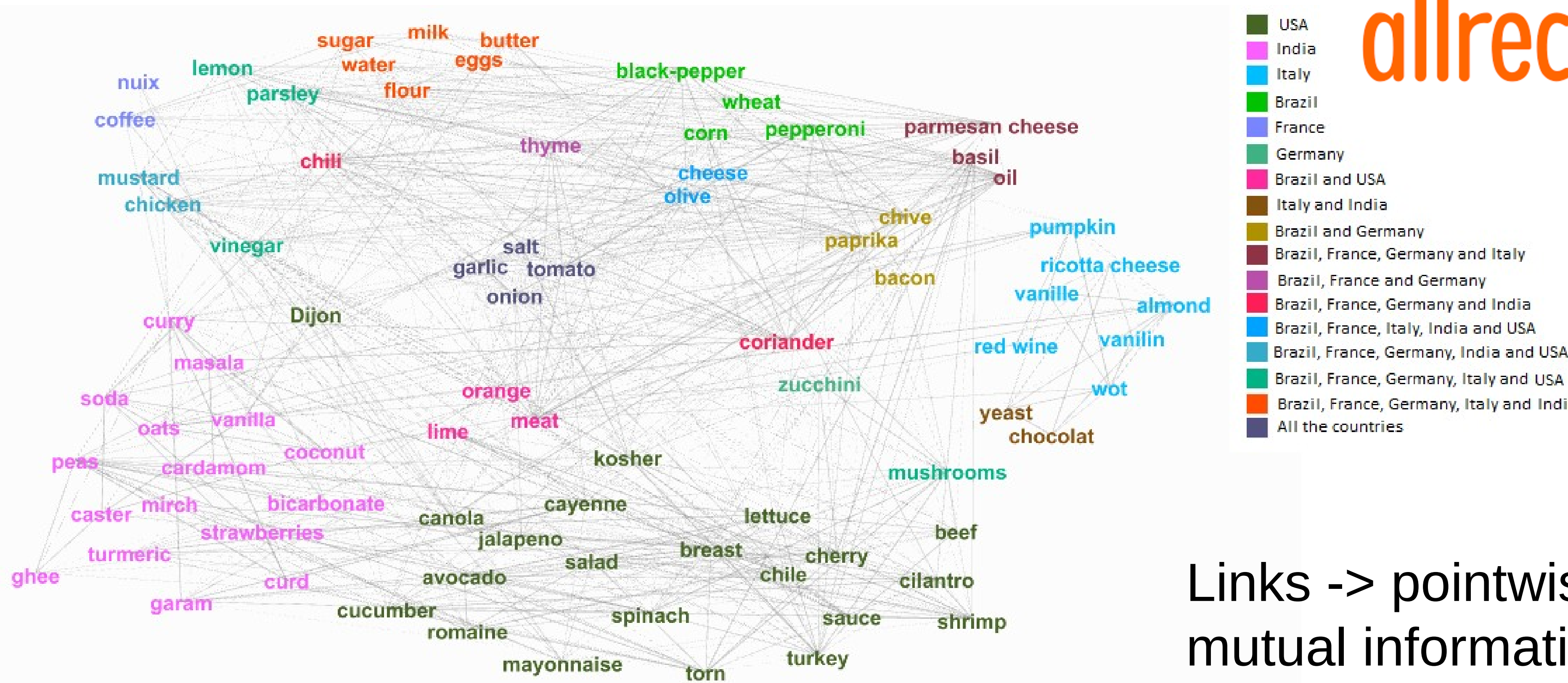
CoUrb 2018, Planning Practice & Research, 2019 (under revision)

Understanding Success

Better understand the success recipes around the world

(New recommendation systems)

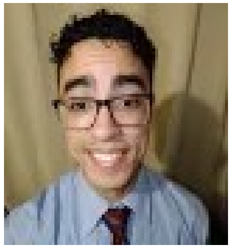
Juliana Viscenheski



Links -> pointwise mutual information

From Pizza to Curry: Preferences for Recipes Around the World
Webmedia 2019

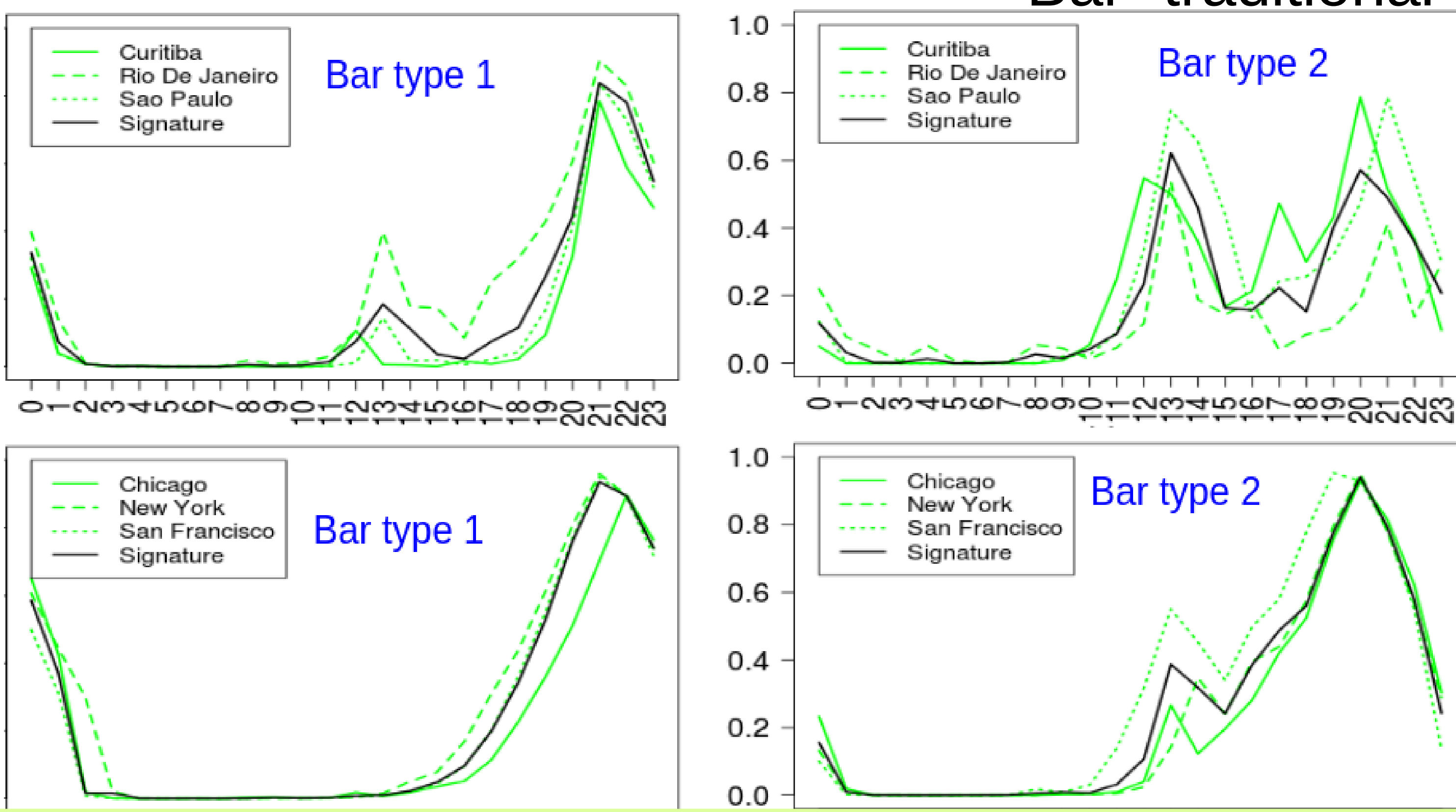
Business Functioning Dynamics

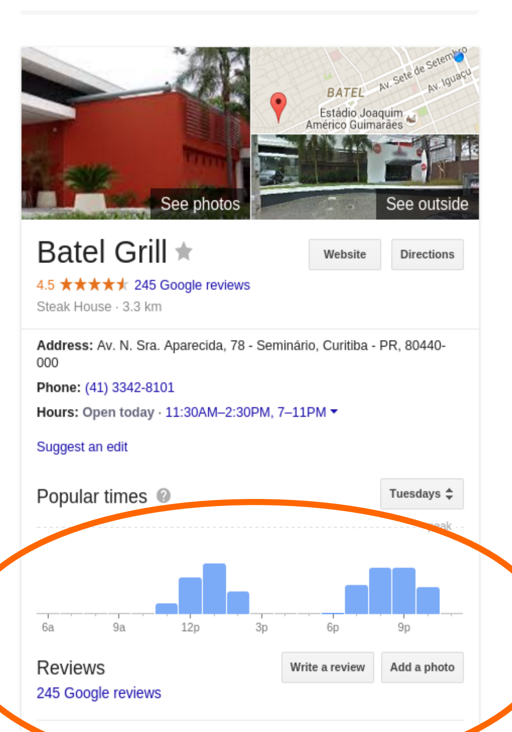


Leonardo de Assis

Popularity time series of places is an important descriptor ("signature" of a place)

Bar "traditional" Bar "nightclub"





Extraction and Exploration of Business Categories Signature
VLDB Workshops, 2018