

# HistorIA

Large historical databases

Interpretability of results

Autonomy of historians with their data

Christophe Prieur & Jean-Daniel Fekete  
Télécom ParisTech & Inria

# Project

historians gather large [for them] amounts of data

- Analyze and explore
  - structural measures
  - clustering
  - network motifs
- But also
  - clean up data
  - visualize

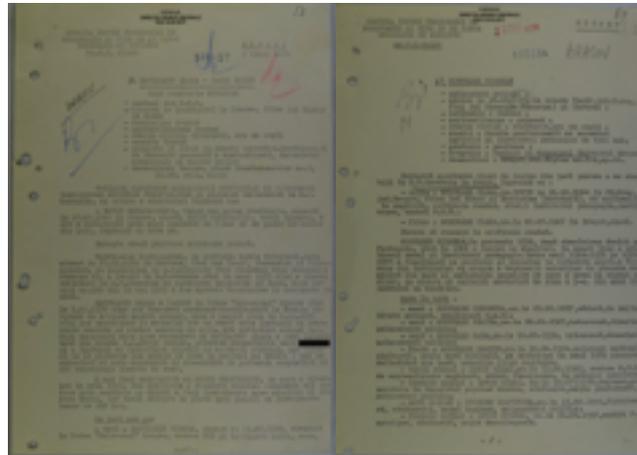
4923) 4-11-1809: Dn. Dionisio BOEDO y LAGO, natural de la ciudad de la Isla de Cereña, hijo legítimo de Dn. Manuel Boedo y de Da. Juana Lago, con  
Da. Catalina PORCEL de PERALTA, natural de ésta, hija legítima de  
Dn. Vicente Porcel de Peralta y de Da. Juana de los Santos. Ts.: Dn.  
Cayetano Gómez y Da. Bernarda Pérez. (f. 23v.).

4924) 11-11-1809: Dn. José PRIMO de RIVERA, Capitán de Fragata gran-  
de de la Real Armada, natural de Aljiciras, hijo legítimo del Se-  
ñor Dn. Brigadier Dn. Joaquín Primo de Rivera, Coronel de Artillería y de la  
Señora Da. Antonia Ortiz de Pinedo Ansochay, con Da. Juana Nemes-  
pucena María de SOBREMONTÉ, natural de la ciudad de Córdoba del  
Virreinato Tucumán, hija legítima del Excelentísimo Señor Dn. Rafael, Marqués  
de Sobremonte, Brigadier de los Reales Ejércitos y ex Virrey de estas  
Provincias y de la Excelentísima Señora Da. Juana María de Larra-  
rañabal y de la Quintana. Ts.: el padre de la contrayente y Da. Juana Ne-  
mescena Larrañabal. (f. 24).

14 000 wedding certificates in Buenos Aires (Ch. Prieur + Zacarias Moutoukias)

```
<ACTE id="13111">13132</date>12-5-1861</date>: <epoux  
id="87"><pre>Venciano</pre> <nom>PARELLI</nom>, de  
<age>28</age> años, natural del <naisance-lieu>Uruguay</naisance-lieu>,  
hijo legítimo de <pers id="89"> <pre>Pedro</pre> <nom>Pedro</nom>  
<nom>Parelli</nom> </pers>, y de <ems id="90"> <pre>Gregoria</pre>  
</nom> <nom>Burgués</nom> </pers>, naturales del  
<naisance-lieu>Uruguay</naisance-lieu>, con <epouse id="88"> <pre>Rosario</pre>  
<nom>SOZA</nom>, de 28 años, natural de <naisance-lieu>Buenos  
Aires</naisance-lieu>, hija legítima de <pers id="91"> <pre>Juan</pre>  
</nom> <nom>Soza</nom> </pers>, natural de  
<naisance-lieu>Portugal</naisance-lieu>, y de <ems id="92"> <pre>Margarita</pre>  
</nom> <nom>Margarita</nom> </pers>, natural de  
<naisance-lieu>Buenos Aires</naisance-lieu> </epouse>. Ts.:  
<temoin> <temoin id="93"> <pre>Agustín</pre> </nom>  
<nom>Robert</nom> </temoin>, y <temoin id="92"> <pre>Margarita</pre>  
</nom> <nom>Margarita</nom> </temoin>. (f.  
42.</ACTE>
```

<http://buenosaires.telecom-paristech.fr/>



Data în țară :  
- soac : DOMINIC GHEORGHE, nr. la 09.06.1897, văduv, de națio-  
nitate germană, pensionar C.A.P.;  
- soră : DOMINIC NEGINA, nr. la 02.01.1907, tricoțesă, divorțată,  
selecționată politic;  
- frate : DOMINIC SARA, nr. la 18.06.1930, tricoțesă, divorțată,  
selecționată politic;  
- frate : DOMINIC MARTIN, nr. la 20.04.1934, medic veterinar,  
căsătorit, soția sora medică. Am solicitat în anul 1970 plecarea  
definitivă din țară, cererile fiind respinse;  
- tatăl soției : IFFIU NIKAL, nr. la 13.06.1905, membru P.C.R.,  
de naționalitate maghiară, văduv, funcționar, în prezent pensionar;  
- fratele soției : IFFIU NIKAL, nr. la 26.07.1930, inginer, șef  
serviciu la Comitetul popular Credea, căsătorit, soția profesoră,  
selecționată politic;  
- sora soției : DOMINIC MĂCĂRINĂ, nr. la 13.02.1915, funcționa-  
ră, căsătorită, soțul inginer, selecționată politic;  
- fratele soției : IFFIU ȘTEFAN, nr. la 26.04.1939, membru P.C.R.  
netrișor, căsătorit, soția dactilografă.

200 000 forms of the Romanian "Securitate" 1950-1989 (Dana Diminescu)

# Problems

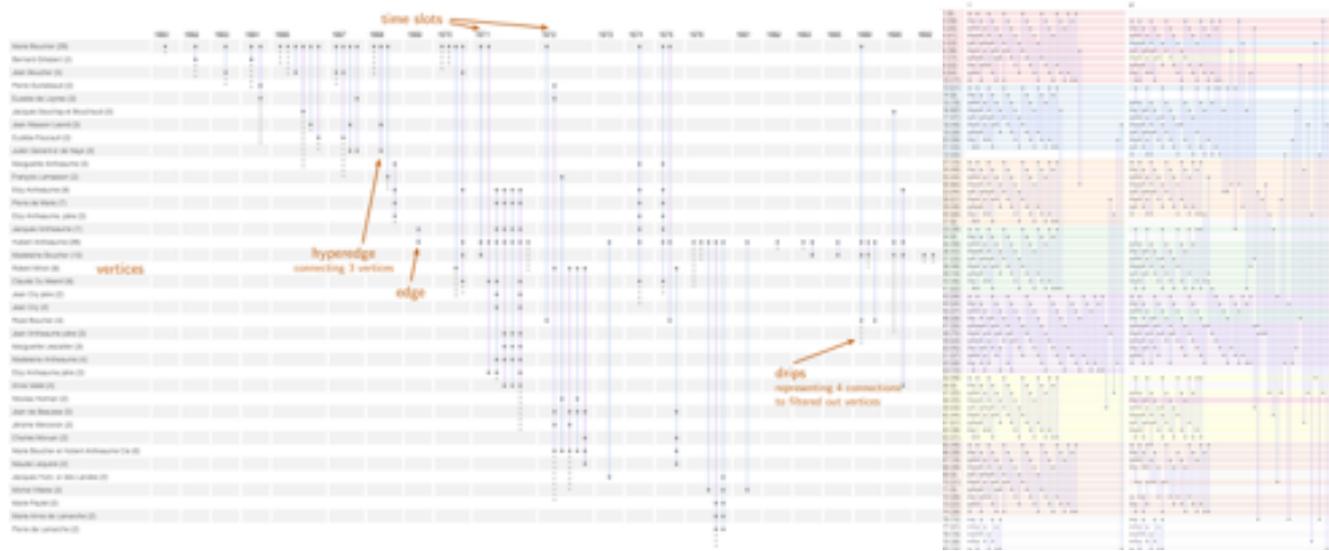
## Scaling up historical data

- Clustering
- Motifs

## Distrust from historians

- « blackbox » algorithms never produce interpretable results
- attributes must be straightforward
  - visualization
  - trial and error

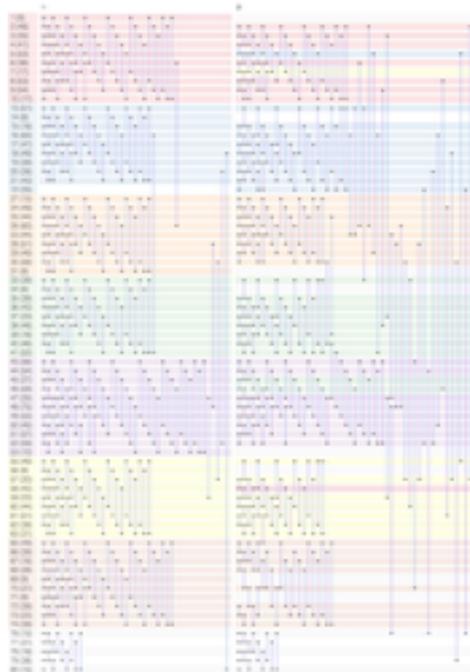
**Data IA challenges :** — From data to knowledge / decisions  
— Transparency and digital trust



# Clustering

Include attributes to produce interpretable groups

- Wedding witnesses
- Place of birth
- Cousins
- Time intervals
- Distance to key individuals

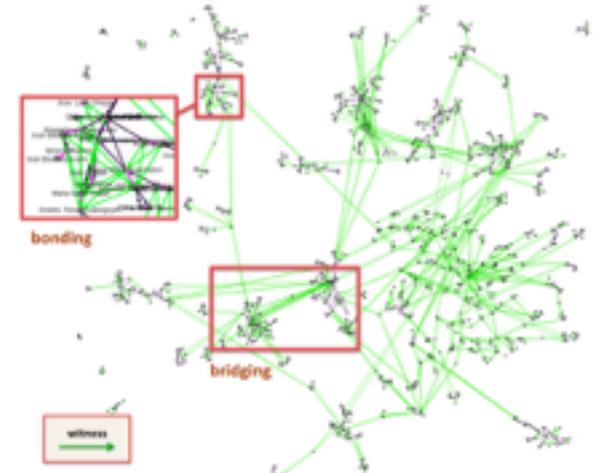


# Motifs

Categorization of elementary combinatorial structures

« Egocentred » networks:  
relational neighborhood of each individual in the corpus

Scaling-up, ahead of state of the art



# Project

Empowering historians with larger and larger data  
(~10 000 individuals, towards millions)

- Platform
  - Ingestion of data
  - Visualization, exploration
  - *Community detection, statistical measures, motifs*
  - *Validation of models by the historians*
  - *Interpretable methods*
- Already there
  - Browsing platform for Buenos Aires database  
<http://buenosaires.telecom-paristech.fr>
  - Vistorian : <https://github.com/networkcube/networkcube>
    - Several visual representations for dynamical networks
  - Extension "back end" Python / NetworkX / ProgressiVis

## • Telecom ParisTech

### Social data science

- Christophe Prieur
- Dana Diminescu

## • Inria Aviz

### dataviz, computer-human interfaces

- Jean-Daniel Fekete

# Team

## • Lied, Univ. Paris Diderot

### History & Complex systems

- Zacarias Moutoukias

## • LaDeHiS, EHESS

### History

- Pascal Cristofoli
- Nicole Dufournaud