PREDICTIONS AND MODELS

Pierre-Alexandre Balland

The Pitch!

- The final exam will reward your creative mind

The Pitch!

- The final exam will reward your creative mind
- Use data science to solve a **real** world problem (no geography needed)

The Pitch!

- The final exam will reward your creative mind
- Use data science to solve a **real** world problem (no geography needed)
- It can be a business, environmental, social problem

The Pitch!

- The final exam will reward your creative mind
- Use data science to solve a **real** world problem (no geography needed)
- It can be a business, environmental, social problem
- You can go beyond this class

The Pitch!

- The final exam will reward your creative mind
- Use data science to solve a **real** world problem (no geography needed)
- It can be a business, environmental, social problem
- You can go beyond this class
- You need to create a pitch deck and convince us that this is a good idea (design + presentation skills)

The Pitch!

- The final exam will reward your creative mind
- Use data science to solve a **real** world problem (no geography needed)
- It can be a business, environmental, social problem
- You can go beyond this class
- You need to create a pitch deck and convince us that this is a good idea (design + presentation skills)
- Groups of 3 students max!

The Pitch!

- The final exam will reward your creative mind
- Use data science to solve a **real** world problem (no geography needed)
- It can be a business, environmental, social problem
- You can go beyond this class
- You need to create a pitch deck and convince us that this is a good idea (design + presentation skills)
- Groups of 3 students max!
- 10 days to prepare your pitch (deadline on November 4)

The Pitch!

- The final exam will reward your creative mind
- Use data science to solve a **real** world problem (no geography needed)
- It can be a business, environmental, social problem
- You can go beyond this class
- You need to create a pitch deck and convince us that this is a good idea (design + presentation skills)
- Groups of 3 students max!
- 10 days to prepare your pitch (deadline on November 4)
- Delivery: video (5 minutes – 5 slides)

AirBed&Breakfast

Book rooms with locals, rather than hotels.

Problem

Price is an important concern for customers booking travel online.

Hotels leave you disconnected from the city and its culture.

No easy way exists to book a room with a local or become a host.

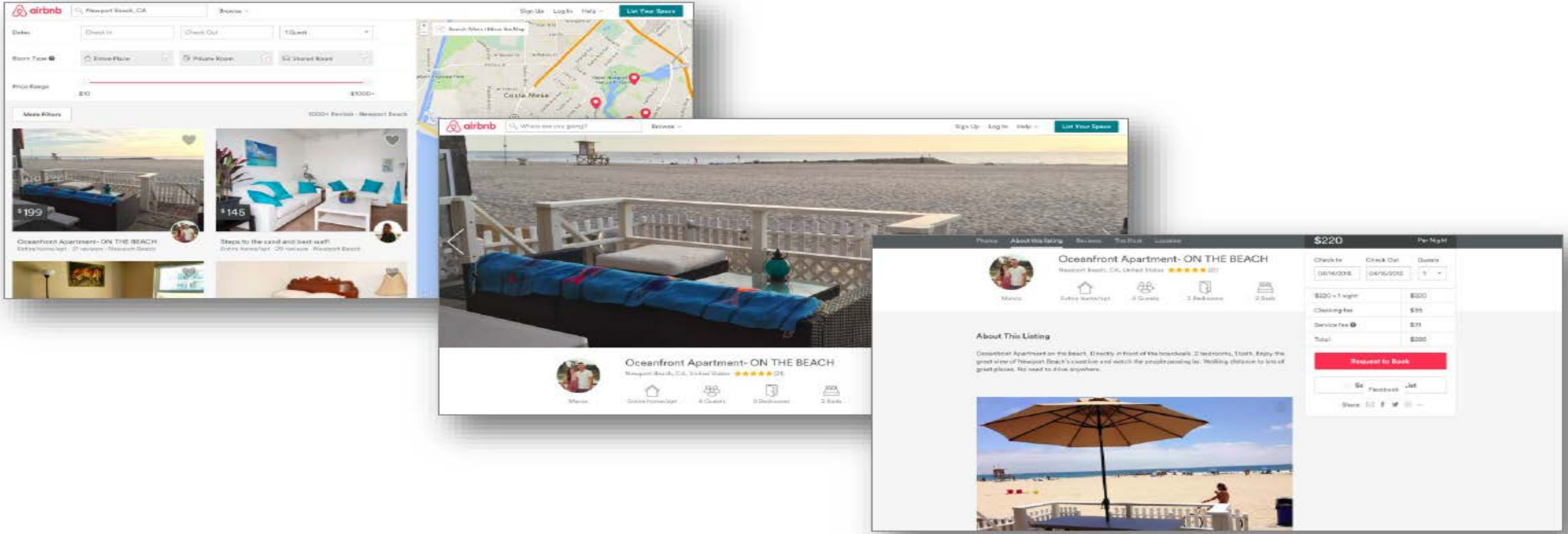
Solution

A web platform where users can rent out their space to host travelers to:

SAVE MONEY	MAKE MONEY	SHARE CULTURE
when traveling	when hosting	local connection to the city

Product

SEARCH BY CITY ⟶ REVIEW LISTINGS ⟶ BOOK IT!



Data science workflow

Import data

Data science workflow

Import data -> Clean

Data science workflow

Import data -> Clean -> Transform

Data science workflow

Import data -> Clean -> Transform -> Visualize

Data science workflow

Import data -> Clean -> Transform -> Visualize -> Model

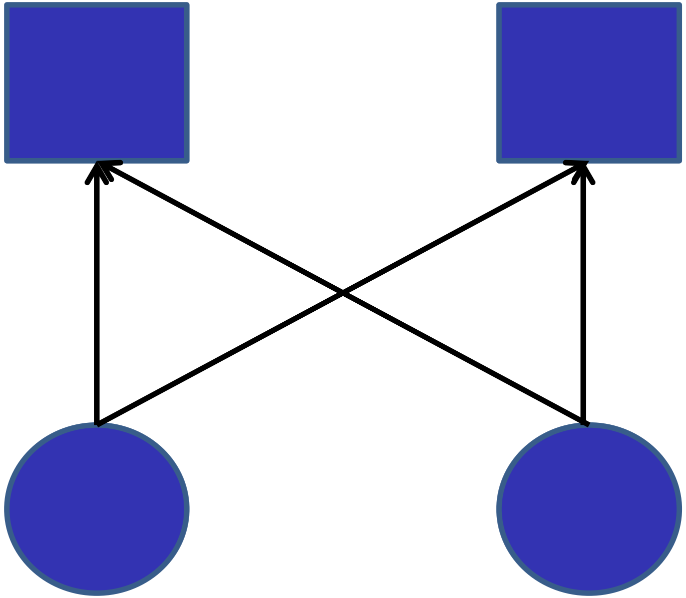
Data science workflow

Import data -> Clean -> Transform -> Visualize -> Model -> Communicate

Key issues

https://www.amazon.science/the-history-of-amazons-recommendation-algorithm

The economy as a two-mode network

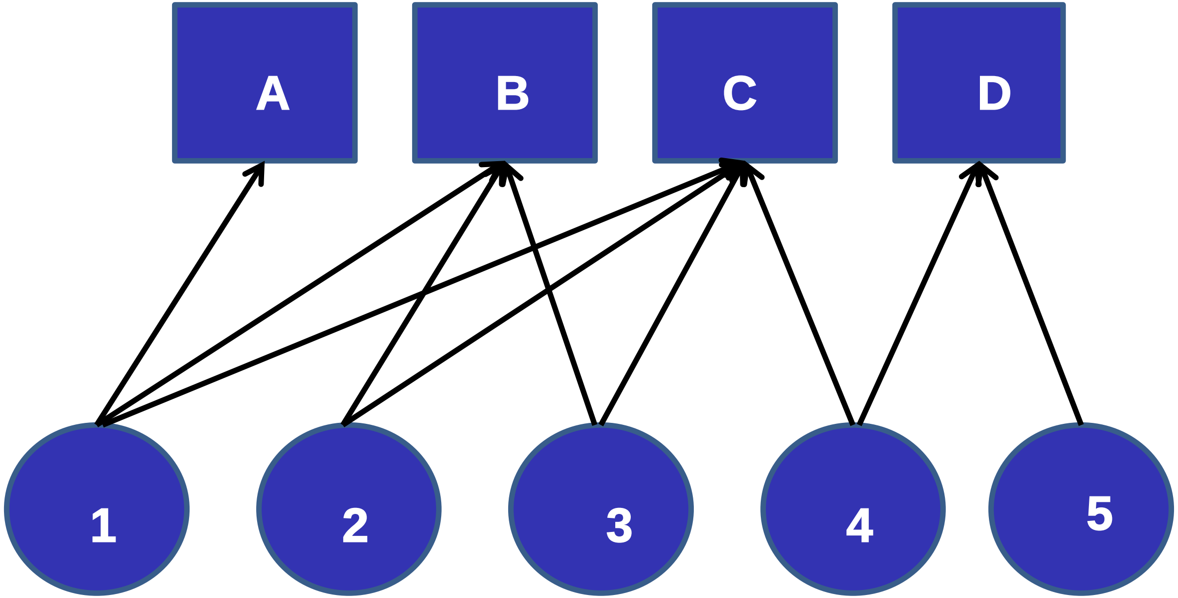


Algorithm’s workflow

Spatial systems as a 2-mode network

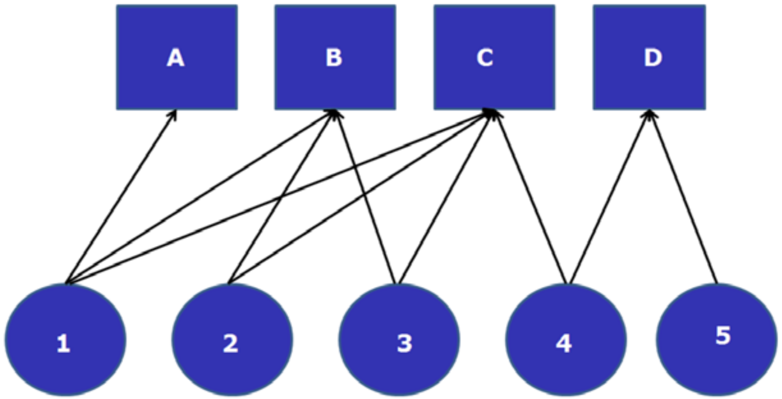
Econ. systems as 2-mode networks

Non-spatial units (economic sectors, jobs, rock band, species, sports…



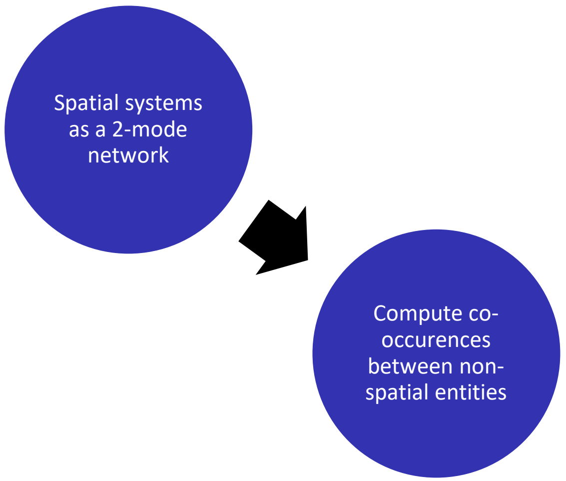
Spatial units (cities, eco-systems, states, neighborhoods…)

2-mode adjacency matrix



$$\begin{array}{c c} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Algorithm’s workflow



Compute co-occurences between non-sp. units

$$
\begin{array}{c c}
 & \begin{array}{cccc} A & B & C & D \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}
\end{array}
$$

How many times B and C co-occur in the same spatial unit?

Compute co-occurences between non-sp. units

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline 1 & & 1 & 1 & \\ 2 & & 1 & 1 & \\ 3 & & 1 & 1 & \\ 4 & & 0 & 1 & \\ 5 & & 0 & 0 & \end{array}$$

How many times B and C co-occur in the same spatial unit?
= in how many spatial unites do B and C co-exist?

Compute co-occurences between non-sp. units

$$
\begin{array}{c}
 \\
1 \\
2 \\
3 \\
4 \\
5
\end{array}
\begin{array}{c}
\begin{array}{cccc} A & B & C & D \end{array} \\
\begin{pmatrix}
 & 1 & 1 & \\
 & 1 & 1 & \\
 & 1 & 1 & \\
 & 0 & 1 & \\
 & 0 & 0 &
\end{pmatrix}
\end{array}
$$

How many times B and C co-occur in the same spatial unit?
= in how many spatial unites do B and C co-exist?
Response = 3 (in 1,2, and 3)

Now using matrix calculus

$$\begin{array}{c|c} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}
\qquad
\begin{array}{c|c} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \begin{pmatrix} \quad & \quad & \quad & \quad \\ & & & \\ & & & \\ & & & \end{pmatrix} \end{array}$$

Find the matrix of co-occurences between all non-spatial units

Transpose the matrix

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Transpose the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Multiply t(M) by M

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Reminder

- For matrix multiplication, the number of **columns** of the first matrix must equal the number of **rows** of the second matrix

- The product of this matrix multiplication will have the same number of **rows** as the first matrix, and the same number of **columns** as the second matrix

- In our case: a 4*5 5*4 matrix will give a ...matrix

$$\begin{pmatrix} 1&0&0&0&0 \\ 1&1&1&0&0 \\ 1&1&1&1&0 \\ 0&0&0&1&1 \end{pmatrix} \times \begin{pmatrix} 1&1&1&0 \\ 0&1&1&0 \\ 0&1&1&0 \\ 0&0&1&1 \\ 0&0&0&1 \end{pmatrix} = \begin{pmatrix} \quad \end{pmatrix}$$

Multiply t(M) by M

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \quad & & & \\ & & & \\ & & & \\ & & & \end{pmatrix}$$

Multiply t(M) by M

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & \\ & & & \\ & & & \\ & & & \end{pmatrix}$$

$1*1+0*0+0*0+0*0+0*0 = 1$

Multiply t(M) by M

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & & \\ & & & \\ & & & \\ & & & \end{pmatrix}$$

*1*1+0*1+0*1+0*0+0*0 = 1*

Multiply t(M) by M

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 3 & 3 & 0 \\ 1 & 3 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

$1*1+1*1+1*1+0*1+0*0 = 3$

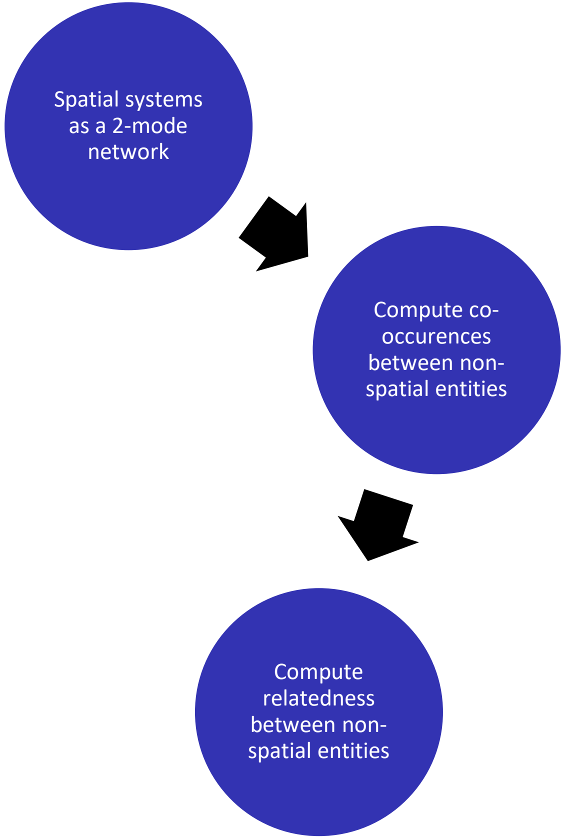
A matrix of co-occurences

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 3 & 3 & 0 \\ 1 & 3 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

A matrix of co-occurences (diag = 0)

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 3 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Algorithm’s workflow



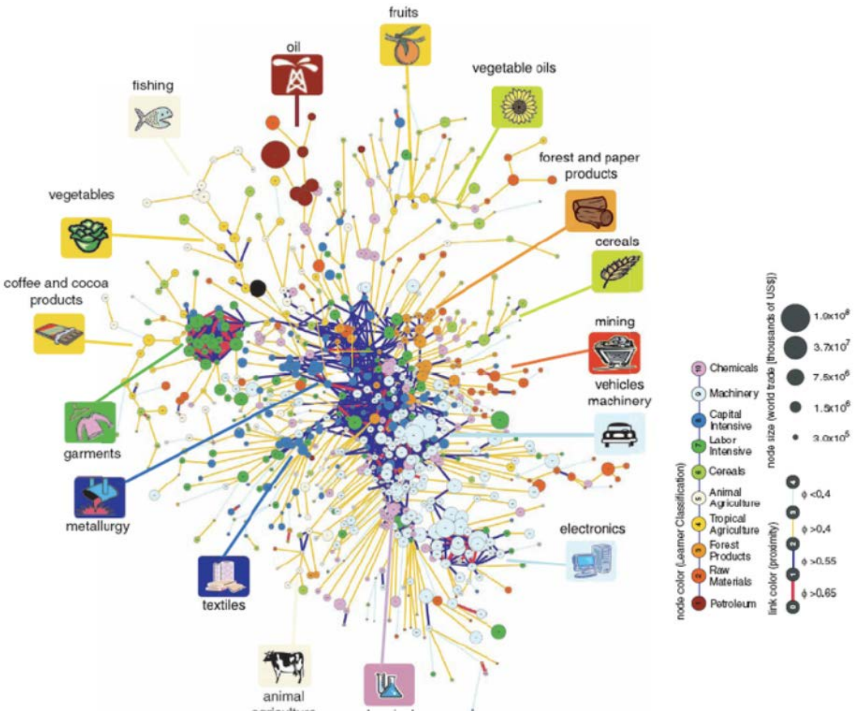
Normalize co-occurences

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 3 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad \frac{observed\ \ co\text{-}occurences}{expected\ \ co\text{-}occurences}$$

Normalizing co-occurences: relatedness

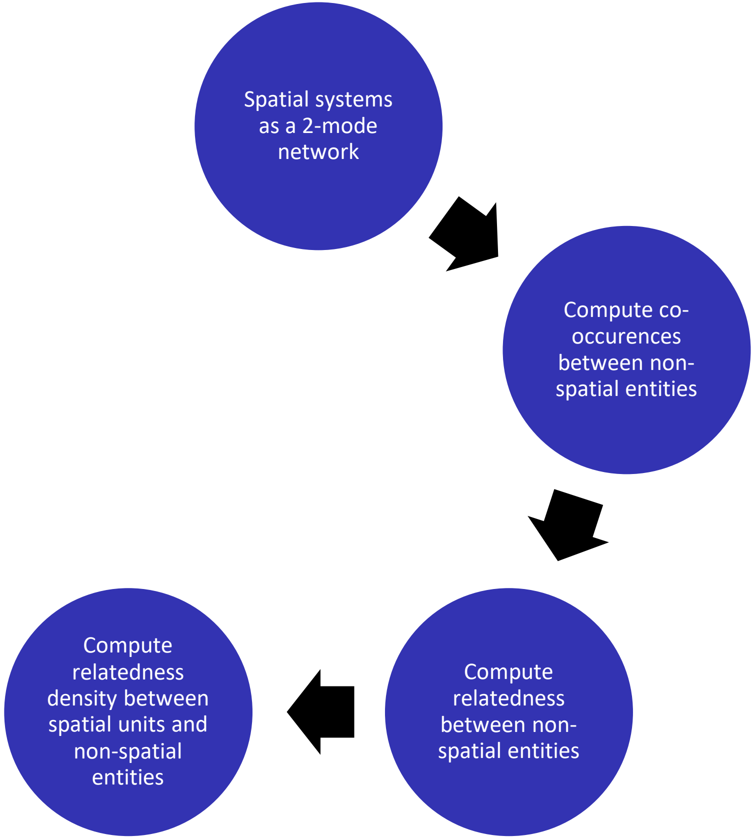
$$\text{if} \quad \frac{\text{observed co-occurences}}{\text{expected co-occurences}} > 1 \dashrightarrow \text{related}$$

Product space & relatedness



Hidalgo et al. (2007), *Science*

Algorithm’s workflow



The density of related technologies

$$D_{i,c,t} = \frac{\sum_i x_i \varphi_{ij}}{\sum_i \varphi_{ij}} \times 100$$

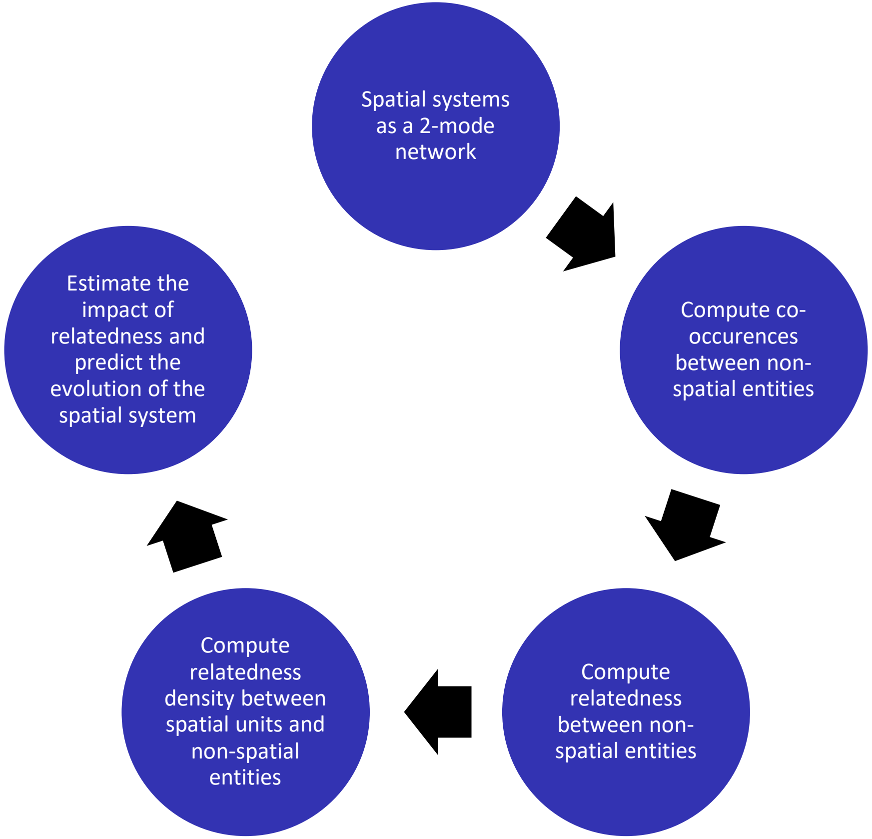
City (MSA)	Tech. class (3 digits)	Density (%)
New-York	428	10
New-York	524	100
Los Angeles	428	80
Los Angeles	524	0
…	…	…

*The **Density Index** measures the relatedness of a new technology to the pre-existing set of technologies produced in this particular city.*

Relatedness density

$$
\begin{array}{c}
 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5
\end{array}
\begin{array}{c}
\begin{array}{cccc} A & B & C & D \end{array} \\
\begin{pmatrix}
100 & 100 & 80 & 100 \\
100 & 75 & 60 & 100 \\
100 & 75 & 60 & 100 \\
50 & 75 & 20 & 100 \\
0 & 0 & 20 & 0
\end{pmatrix}
\end{array}
$$

Algorithm’s workflow



$$\boxed{Entry_{i,c,t} = \beta_1 Density_{i,c,t-1} + \beta_2 City_{c,t-1} + \beta_3 Techno_{i,t-1} + \phi_c + \psi_i + \alpha_t + \varepsilon_{i,c,t}}$$

***Entry$_{i,c,t}$* = 1** *if a technology i that did not belong to the portfolio of the city c in time t-1 enters its technology space in time t.*

Dependent variable is: $Entry_t$	Model 1 Rel. density	Model 2 City variables	Model 3 Tech. variables	Model 4 Full model	Model 5 Full model (F.E.)
Relatedness $density_{t-1}$	0.00515979**			0.00373407**	0.00271463**
	(0.00012770)			(0.00014135)	(0.00016884)
Log $(Employment)_{t-1}$		0.04934166**		0.03611889**	0.04633250**
		(0.00286818)		(0.00247147)	(0.00782869)
Population $density_{t-1}$		0.00001106		0.00002520**	−0.00021341**
		(0.00000997)		(0.00000843)	(0.00003836)
Inventive $capacity_{t-1}$		0.07718815**		0.03883926**	−0.08487966**
		(0.01294204)		(0.0078352020)	(0.01505564)
Tech. $Specialization_{t-1}$		−0.00089296**		−0.00047160**	0.00005120
		(0.00011548)		(0.00009315)	(0.00011022)
MSA growth $rate_{t-1}$		0.04443962**		0.04032813**	0.00865397**
		(0.00355534)		(0.00353667)	(0.00298386)
Log (Income per $employee)_{t-1}$		−0.07584685**		−0.10127439**	0.00368879
		(0.00441610)		(0.00538561)	(0.01663469)
Log (Nb. $Inventors)_{t-1}$			0.02658895**	0.02324554**	0.00159990
			(0.00197752)	(0.00183672)	(0.00246612)
Tech. $concentration_{t-1}$			−0.00102840**	−0.00010693	0.00041990 *
			(0.00014936)	(0.00011541)	(0.00016760)
Date $established_{t-1}$			−0.00056684**	−0.00042520**	−0.00330620**
			(0.00007012)	(0.00005456)	(0.00017699)
Tech. growth $rate_{t-1}$			0.01423964**	0.02183910**	0.01141729**
			(0.00233334)	(0.00285492)	(0.00260757)
Constant	0.09258502**	0.09296771**	0.09019069**	0.08909252**	0.11108572**
	(0.00194271)	(0.00378306)	(0.00398429)	(0.00183778)	(0.01040890)
City F.E.	No	No	No	No	Yes
Technology F.E.	No	No	No	No	Yes
Period F.E.	No	No	No	No	Yes
R^2	0.11	0.04	0.02	0.13	0.16
N	748,458	653,660	656,618	572,550	572,550