UNIVERSITY OF BIRMINGHAM

Interdependence and predictability of human mobility and social interactions

M. De Domenico[1]

[1]School of Computer Science, University of Birmingham, UK

Mathematics of Networks
University of Warwick, 20^{th} July 2012

My background...

UNIVERSITY OF BIRMINGHAM

I come from Italy...

UNIVERSITÀ di CATANIA

SCUOLA SUPERIORE di CATANIA
MEDITERRANEAN UNIVERSITY CENTER

Astroparticle physics
Cosmic rays

Complex systems
Nonlinear time series analysis
Chaos theory
Information theory
Extreme value theory

In UK since March...

UNIVERSITY OF BIRMINGHAM

Complex systems
Network science
Nonlinear time series analysis

Outline of this talk

UNIVERSITY OF BIRMINGHAM

Part I. Interdependence

- Networks of time series
- Correlation measures

Part II. Predictability

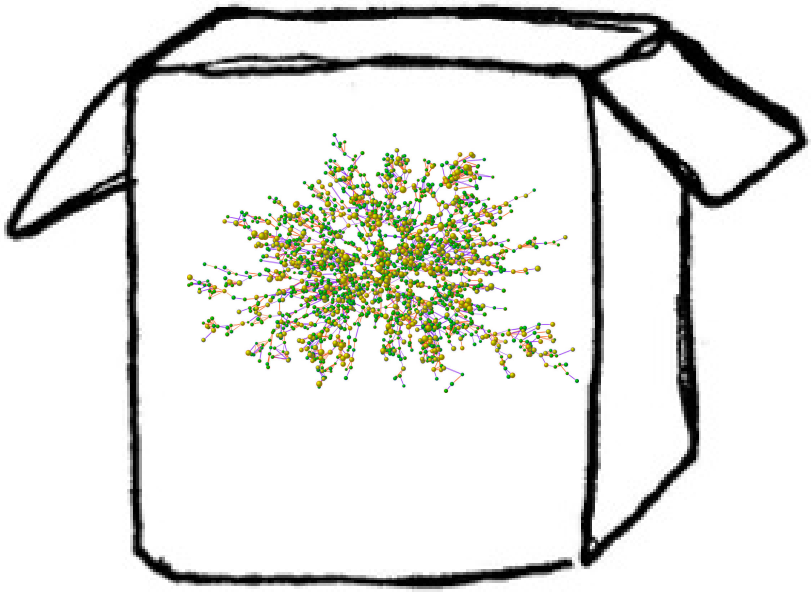
- "Embeddology"
- (Multivariate) Nonlinear Predictor

Part III. Application to human mobility

- Is it possible to *predict* human movements?
- The Nokia Mobile Data Challange

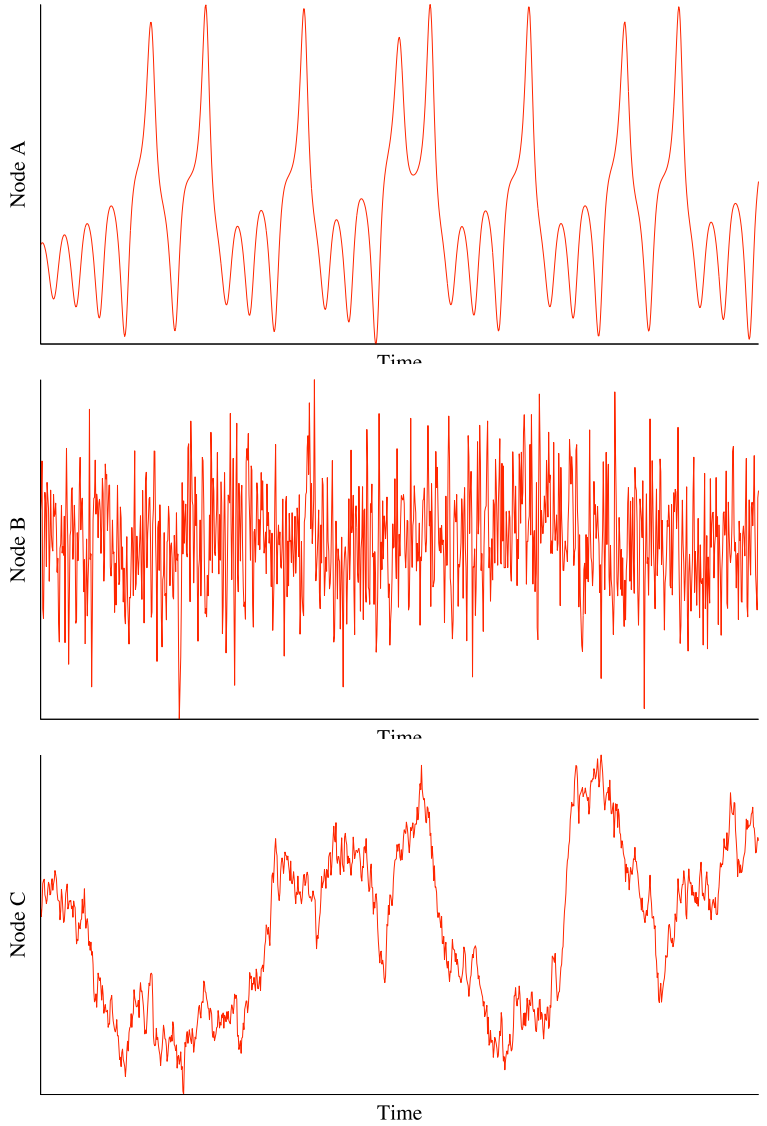
UNIVERSITY OF BIRMINGHAM

(Unknown) network of dynamical systems



Real-world

Observations! $\Longrightarrow$

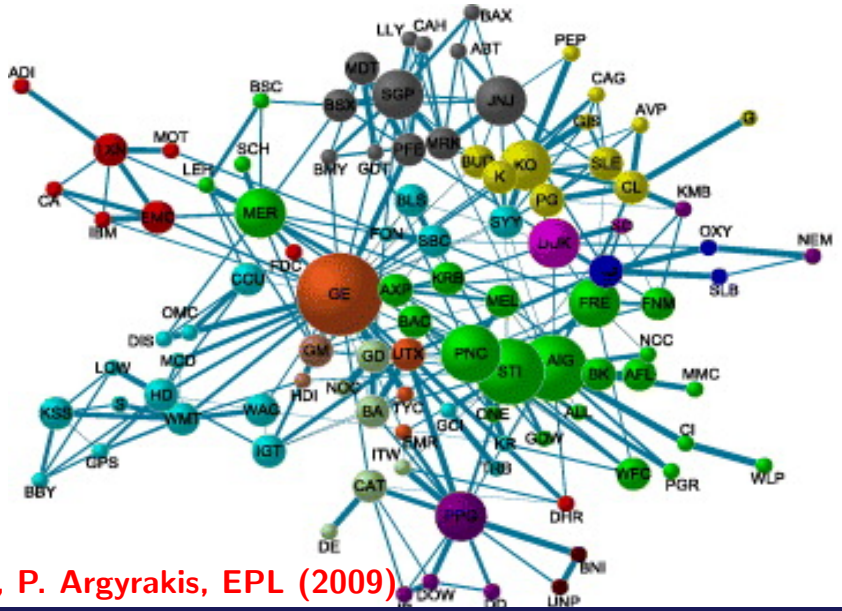


Goal: build the network of correlations among stocks from several univariate time series of stock prices

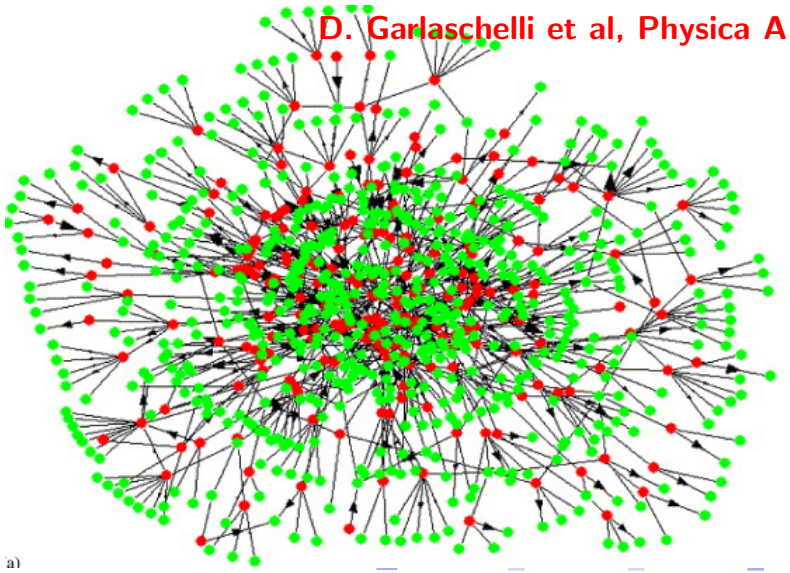
Correlation coefficient:

$$\rho_{ij} = \frac{\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle}{\sqrt{[\langle x_i^2 \rangle - \langle x_i \rangle^2][\langle x_j^2 \rangle - \langle x_j \rangle^2]}}$$

Distance matrix elements given by $D_{ij} = \sqrt{2(1-\rho_{ij})}$, to build the network of correlations (R.N. Mantegna, Eur. Phys. J. B (1999))



A. Garas, P. Argyrakis, EPL (2009)

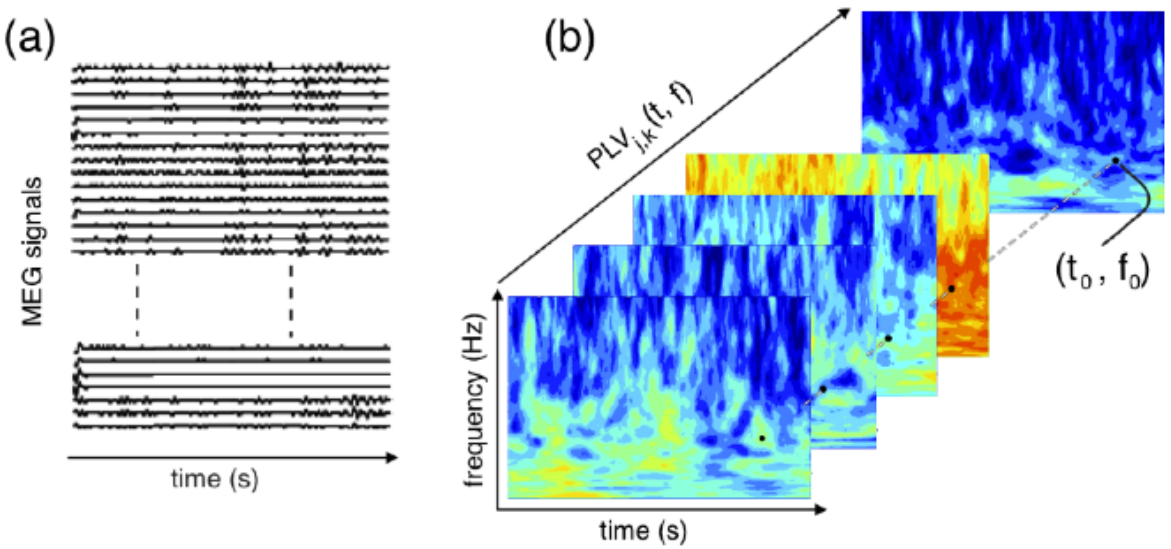


D. Garlaschelli et al, Physica A (2005)

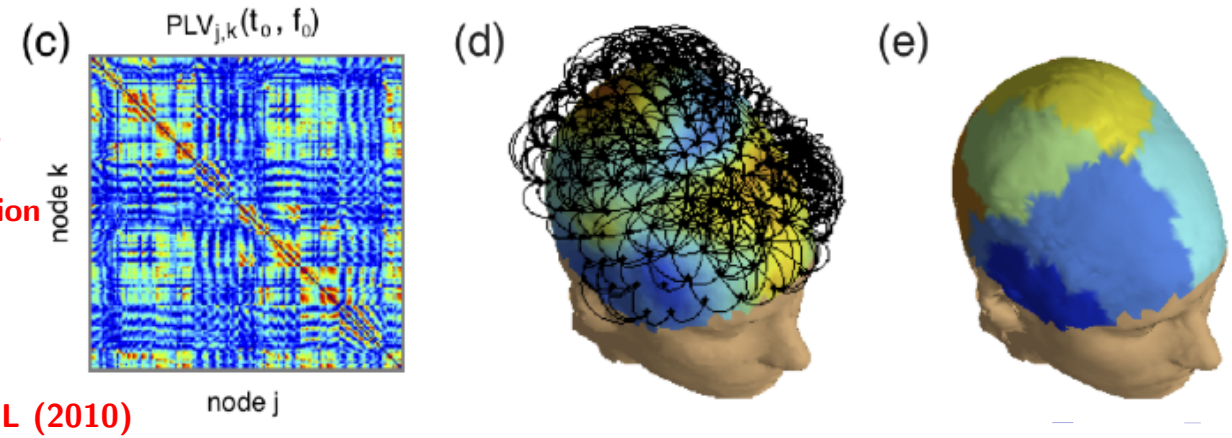
Network from time series: brain

Goal: investigate brain connectivity to detect unhealthy (epileptic) subjects

Time series



Phase-Locking Values as measure of correlation/synchronization

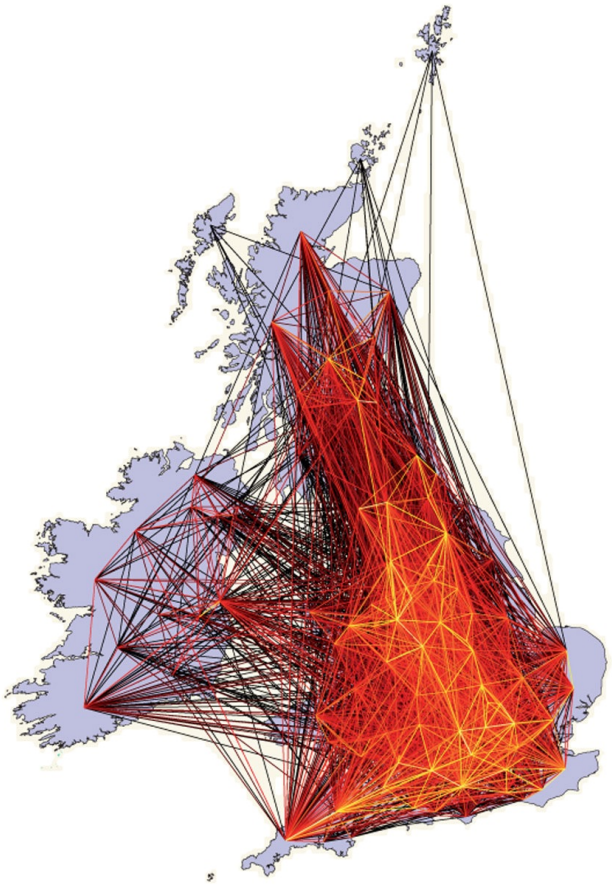


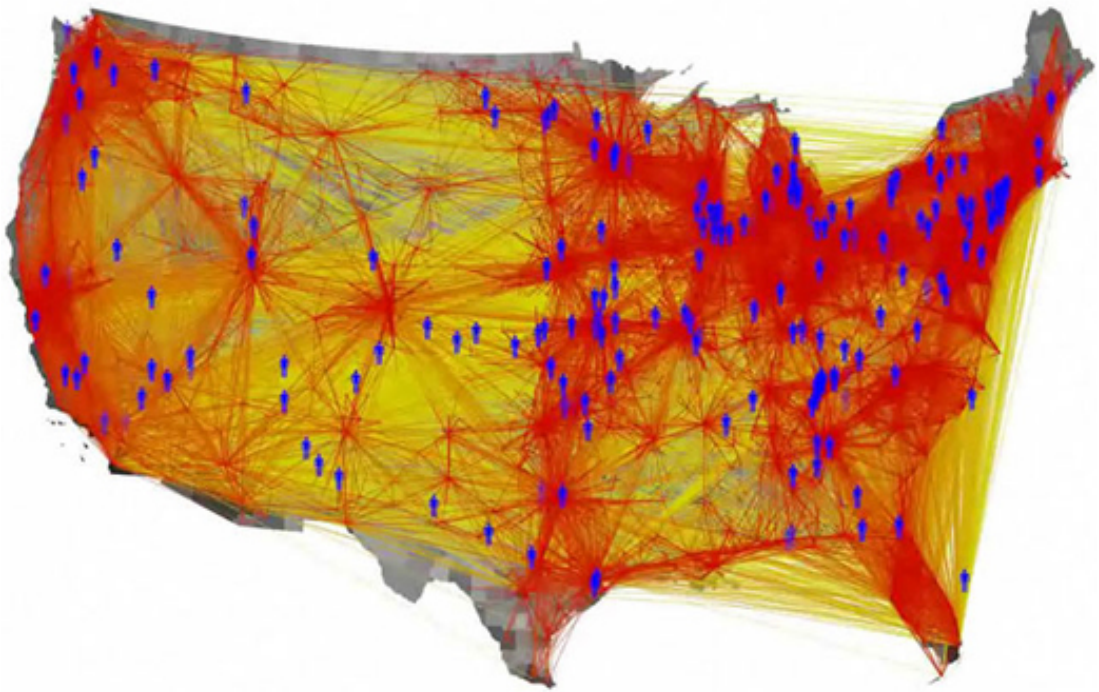
M. Chavez et al, PRL (2010)

What about human movements?

UNIVERSITY OF BIRMINGHAM

Goal: investigate people interactions from their mobility patterns for geo-prediction purposes





D. Brockmann, Nature (2012)

Image credit: Christian Thiemann and Daniel Grady, Northwestern University

What about human interactions?

UNIVERSITY OF BIRMINGHAM

- Deal with 2-variate (or 3-variate) time series measurements, e.g., GPS readings, for each individual
- **Drawbacks** of standard cross-correlation:
  - accounts **only** for linear correlations $\longrightarrow$ unable to capture nonlinear features
  - is not trivial to extend to multivariate measurements

Hence, we propose information theoretical measures to capture **correlations**. Advantages:

- Based on the rather general concept of **information**
- Estimated from probability density
- Do not make assumptions on the underlying dynamics
- Able to capture nonlinear correlations

What about human interactions?

UNIVERSITY OF BIRMINGHAM

- Deal with 2-variate (or 3-variate) time series measurements, e.g., GPS readings, for each individual
- **Drawbacks** of standard cross-correlation:
  - accounts **only** for linear correlations $\longrightarrow$ unable to capture nonlinear features
  - is not trivial to extend to multivariate measurements

Hence, we propose information theoretical measures to capture **correlations**. Advantages:

- Based on the rather general concept of **information**
- Estimated from probability density
- Do not make assumptions on the underlying dynamics
- Able to capture nonlinear correlations

UNIVERSITY OF BIRMINGHAM

- Let $\mathbf{X}$ be a multivariate stochastic variable (e.g., GPS reading)
- $P_{\mathbf{X}}(\mathbf{x})$ is the *true* probability density function (PDF)
- $Q_{\mathbf{X}}(\mathbf{x})$ is some *approximate model* describing outcomes of $\mathbf{X}$

Question: What is the price to pay for the incompleteness of our model to describe the real underlying distribution? Equivalently, what is the gain of information about the data if we use our model?

Answer: It is quantified by the Kullback-Leibler divergence (1951, 1959)

$$\mathcal{D}_{KL}(P||Q) = \sum_{\mathbf{x}\in\mathbf{X}} P_{\mathbf{X}}(\mathbf{x}) \log \frac{P_{\mathbf{X}}(\mathbf{x})}{Q_{\mathbf{X}}(\mathbf{x})} = \underbrace{\mathcal{H}(P,Q)}_{\substack{\text{Shannon}\\ \text{cross-entropy}}} - \underbrace{\mathcal{H}(P)}_{\substack{\text{Shannon}\\ \text{entropy}}}$$

- $\mathcal{D}_{KL}(P||Q) \geq 0$ (not bounded above: **undesirable feature**)
- $\mathcal{D}_{KL}(P||Q) = 0 \Leftrightarrow P_{\mathbf{X}}(\mathbf{x}) = Q_{\mathbf{X}}(\mathbf{x})$
- $\mathcal{D}_{KL}(P||Q) \neq \mathcal{D}_{KL}(Q||P)$ (**undesirable feature**)

- Let $\mathbf{X}$ be a multivariate stochastic variable (e.g., GPS reading)
- $P_{\mathbf{X}}(\mathbf{x})$ is the *true* probability density function (PDF)
- $Q_{\mathbf{X}}(\mathbf{x})$ is some *approximate model* describing outcomes of $\mathbf{X}$

Question: What is the price to pay for the incompleteness of our model to describe the real underlying distribution? Equivalently, what is the gain of information about the data if we use our model?

Answer: It is quantified by the Kullback-Leibler divergence (1951, 1959)

$$\mathcal{D}_{KL}(P||Q) = \sum_{\mathbf{x}\in\mathbf{X}} P_{\mathbf{X}}(\mathbf{x}) \log \frac{P_{\mathbf{X}}(\mathbf{x})}{Q_{\mathbf{X}}(\mathbf{x})} = \underbrace{\mathcal{H}(P,Q)}_{\substack{\text{Shannon}\\ \text{cross-entropy}}} - \underbrace{\mathcal{H}(P)}_{\substack{\text{Shannon}\\ \text{entropy}}}$$

- $\mathcal{D}_{KL}(P||Q) \geq 0$ (not bounded above: **undesirable feature**)
- $\mathcal{D}_{KL}(P||Q) = 0 \Leftrightarrow P_{\mathbf{X}}(\mathbf{x}) = Q_{\mathbf{X}}(\mathbf{x})$
- $\mathcal{D}_{KL}(P||Q) \neq \mathcal{D}_{KL}(Q||P)$ (**undesirable feature**)

Possible alternative is the Jensen-Shannon divergence

$$\mathcal{D}_{JS}(P||Q) = \frac{1}{2}\mathcal{D}_{KL}(P||R) + \frac{1}{2}\mathcal{D}_{KL}(Q||R), \quad R = \frac{P+Q}{2}$$

- **Bounded:** $0 \leq \mathcal{D}_{JS}(P||Q) \leq 1$
- $\mathcal{D}_{JS}(P||Q) = 0 \Leftrightarrow P_{\mathbf{X}}(\mathbf{x}) = Q_{\mathbf{X}}(\mathbf{x})$
- **Symmetric:** $\mathcal{D}_{JS}(P||Q) = \mathcal{D}_{JS}(Q||P)$

Application to human mobility

- $\mathbf{X}$ represents the motion of a user on the Earth
- Random samples $\mathbf{x}$ ($\mathbf{y}$) drawn from $\mathbf{X}$ ($\mathbf{Y}$) correspond to geographic coordinates
- The PDF of $\mathbf{x}$ ($\mathbf{y}$) quantifies the **fraction of time spent by the user X (Y) in a particular position**
- Use $\mathcal{D}_{JS}(P_{\mathbf{X}}||P_{\mathbf{Y}})$ and $\mathcal{D}_{KL}(P_{\mathbf{X}}||P_{\mathbf{Y}})$ to quantify the similarity of their mobility patterns

Mutual information is another way of measuring the **correlation**:

$$\mathcal{I}(\mathbf{X},\mathbf{Y}) = \sum_{\mathbf{x}\in\mathbf{X}}\sum_{\mathbf{y}\in\mathbf{Y}} P_{\mathbf{XY}}(\mathbf{x},\mathbf{y}) \log \frac{P_{\mathbf{XY}}(\mathbf{x},\mathbf{y})}{P_{\mathbf{X}}(\mathbf{x})P_{\mathbf{Y}}(\mathbf{y})} = \mathcal{D}_{KL}(P_{\mathbf{XY}}||P_{\mathbf{X}}P_{\mathbf{Y}})$$

where $P_{\mathbf{XY}}$ is the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$.

- Quantifies how much information the variable $\mathbf{Y}$ provides about the variable $\mathbf{X}$
- If $\mathbf{X}$ and $\mathbf{Y}$ are totally uncorrelated: $P_{\mathbf{XY}} = P_{\mathbf{X}}P_{\mathbf{Y}}$ and $\mathcal{I}(\mathbf{X},\mathbf{Y}) = 0$
- Robust estimator of correlation, but suffers from the same undesirable features of Kullback-Leibler divergence

Statistical validation of the method

UNIVERSITY OF BIRMINGHAM

- How to choose the best measure?
- Statistical analysis on controlled data (toy models) required
- **Assumption:** an individual moves *randomly* only on small spatio-temporal scales, but *regularly enough* on larger scales $\Longrightarrow$ he/she is not a random walker, he/she is more likely to be a chaotic one (long-term unpredictable, following complex spatio-temporal patterns)

Simulation setup

- Simulate agents moving on a geographical surface according to some chaotic pattern (10 different chaotic dynamics simulated)
- Make it more realistic: add correlated (pink) noise to patterns
- Tunable parameters: i) **# samples**, ii) **signal-to-noise ratio** (SNR)
- Several random realizations of the same setup: $k = 1, 2, ..., N$

Statistical validation of the method

UNIVERSITY OF BIRMINGHAM

- How to choose the best measure?
- Statistical analysis on controlled data (toy models) required
- **Assumption:** an individual moves *randomly* only on small spatio-temporal scales, but *regularly enough* on larger scales $\Longrightarrow$ he/she is not a random walker, he/she is more likely to be a chaotic one (long-term unpredictable, following complex spatio-temporal patterns)

Simulation setup

- Simulate agents moving on a geographical surface according to some chaotic pattern (10 different chaotic dynamics simulated)
- Make it more realistic: add correlated (pink) noise to patterns
- Tunable parameters: i) **# samples**, ii) **signal-to-noise ratio** (SNR)
- Several random realizations of the same setup: $k = 1, 2, ..., N$

Statistical validation of the method

UNIVERSITY OF BIRMINGHAM

Procedure

1. Consider 20 chaotic agents, fixing number of samples and SNR
2. Generate N random undir. unw. 20×20 adjacency matrices A_k
3. If $a_{ij}^{(k)} = 1$, generate the same chaotic pattern for agents i_k and j_k, add noise (with different seeds) to the traces
4. Use correlation measures to obtain the matrix B_k
5. Use a similarity measure to estimate how much B_k approximates A_k
6. Vary sample length, SNR and repeat from (1)

As a similarity measure, we use the Frobenius norm of $B - A$, normalized to $[0, 1]$, defined by

$$\phi = \frac{1}{n}\sqrt{(B-A)(B-A)^\dagger} = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n} |b_{ij} - a_{ij}|^2}$$

where $n = 20$. If $B = A$ then $\phi = 0$, otherwise $0 < \phi \leq 1$.

Statistical validation of the method

UNIVERSITY OF BIRMINGHAM

Procedure

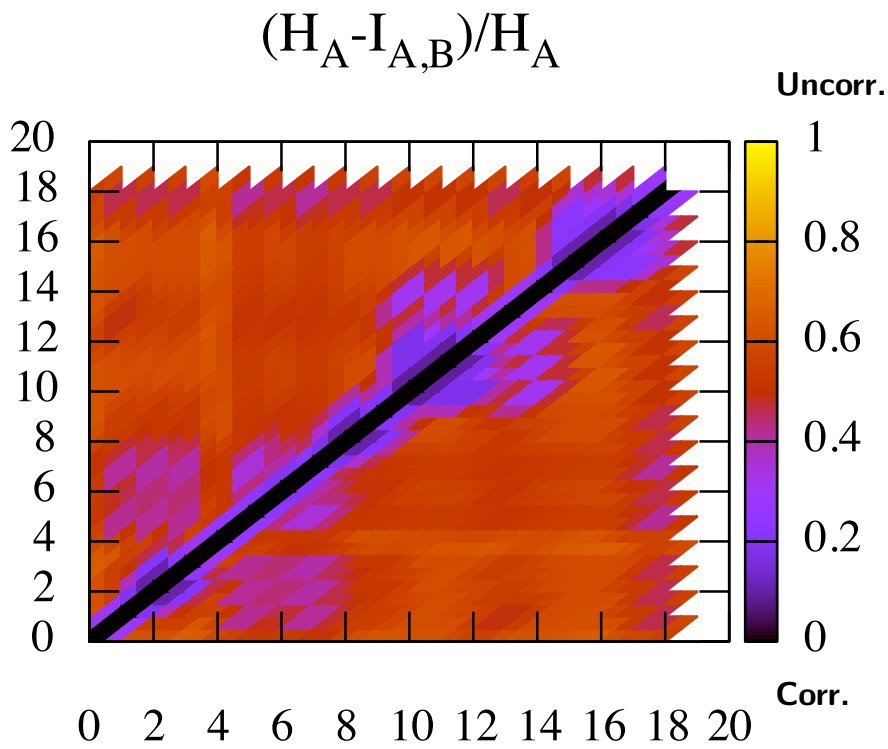
1. Consider 20 chaotic agents, fixing number of samples and SNR
2. Generate N random undir. unw. 20×20 adjacency matrices A_k
3. If $a_{ij}^{(k)} = 1$, generate the same chaotic pattern for agents i_k and j_k, add noise (with different seeds) to the traces
4. Use correlation measures to obtain the matrix B_k
5. Use a similarity measure to estimate how much B_k approximates A_k
6. Vary sample length, SNR and repeat from (1)

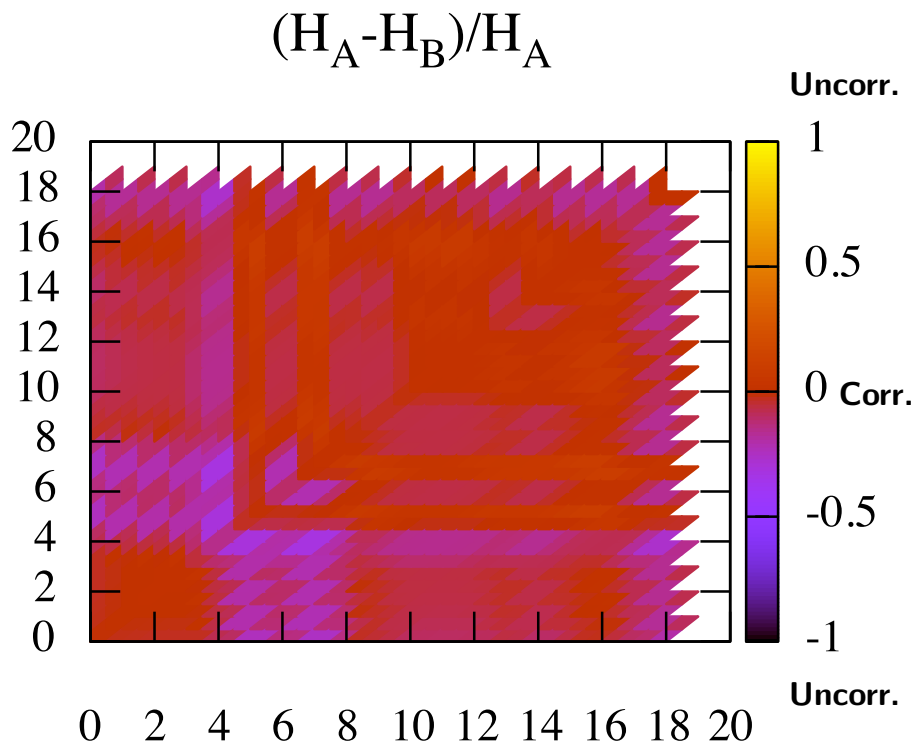
As a similarity measure, we use the Frobenius norm of $B - A$, normalized to $[0, 1]$, defined by

$$\phi = \frac{1}{n}\sqrt{(B-A)(B-A)^\dagger} = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n} |b_{ij} - a_{ij}|^2}$$

where $n = 20$. If $B = A$ then $\phi = 0$, otherwise $0 < \phi \leq 1$.

Correlation matrices



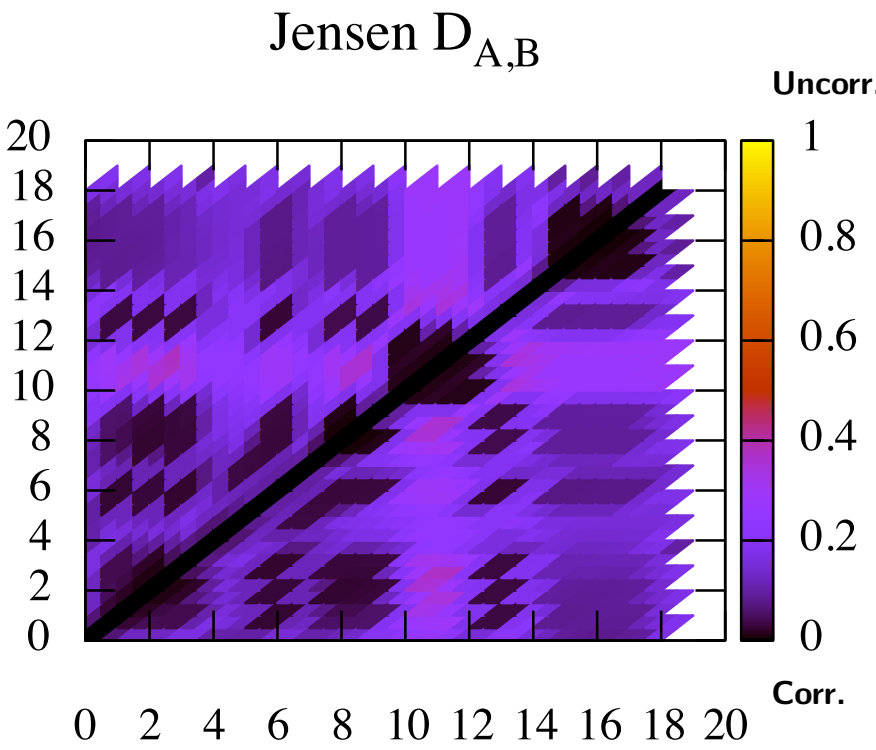


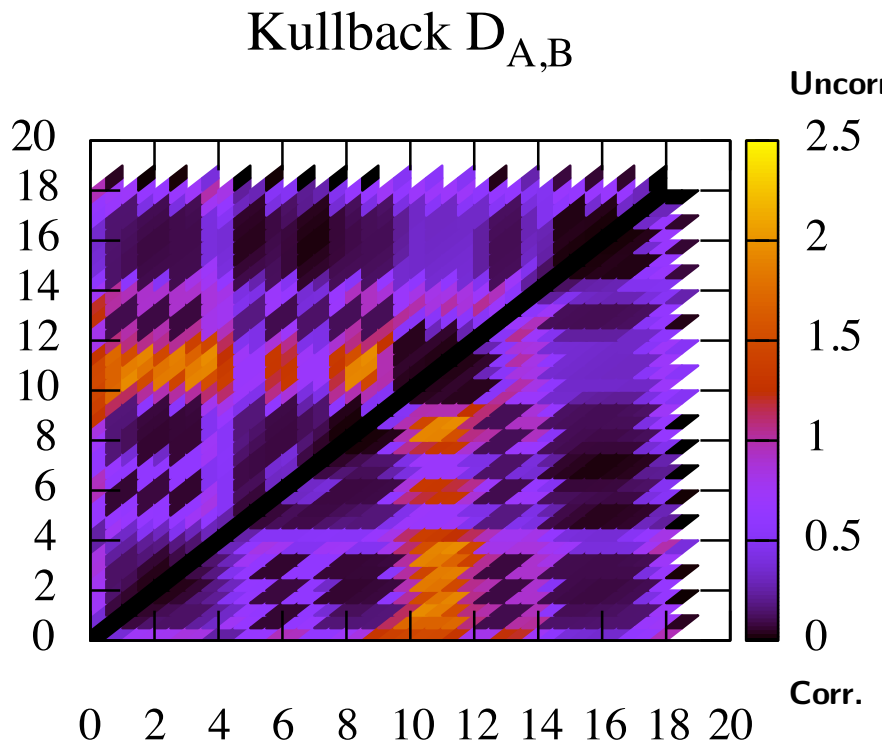
NodeID vs **NodeID**

Inferred correlation matrix $B = (b_{ij}) \in \mathbb{R}^{n \times n}$

$$b_{ij} = \begin{cases} 1 & \text{if } b^{\star}_{ij} \geq (\leq) \text{ thresh} \\ 0 & \text{if } b^{\star}_{ij} < (>) \text{ thresh} \end{cases}$$

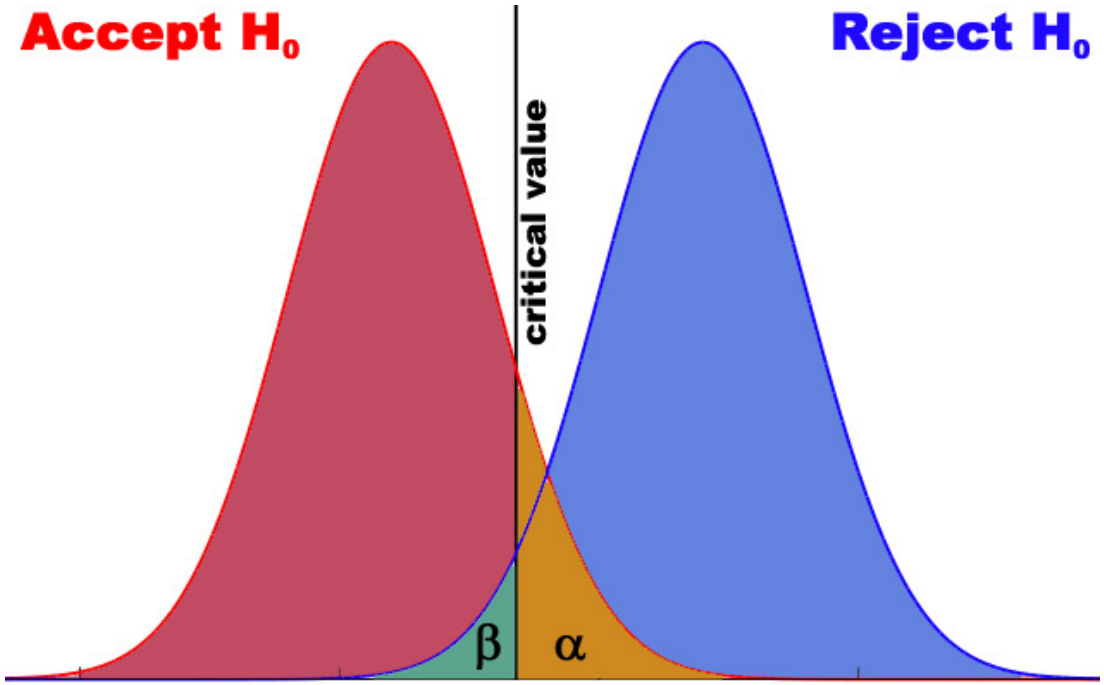
Correlation matrices





NodeID vs **NodeID**

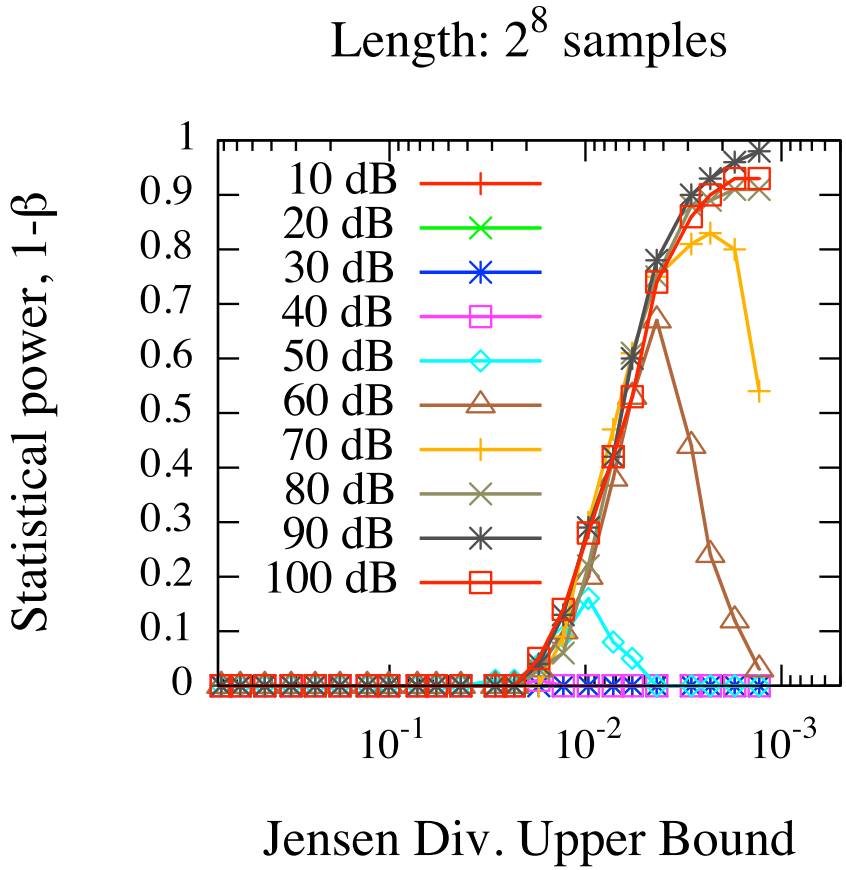
Our numerical studies show **Jensen-Shannon divergence** and **Mutual Information** are the **most promising**

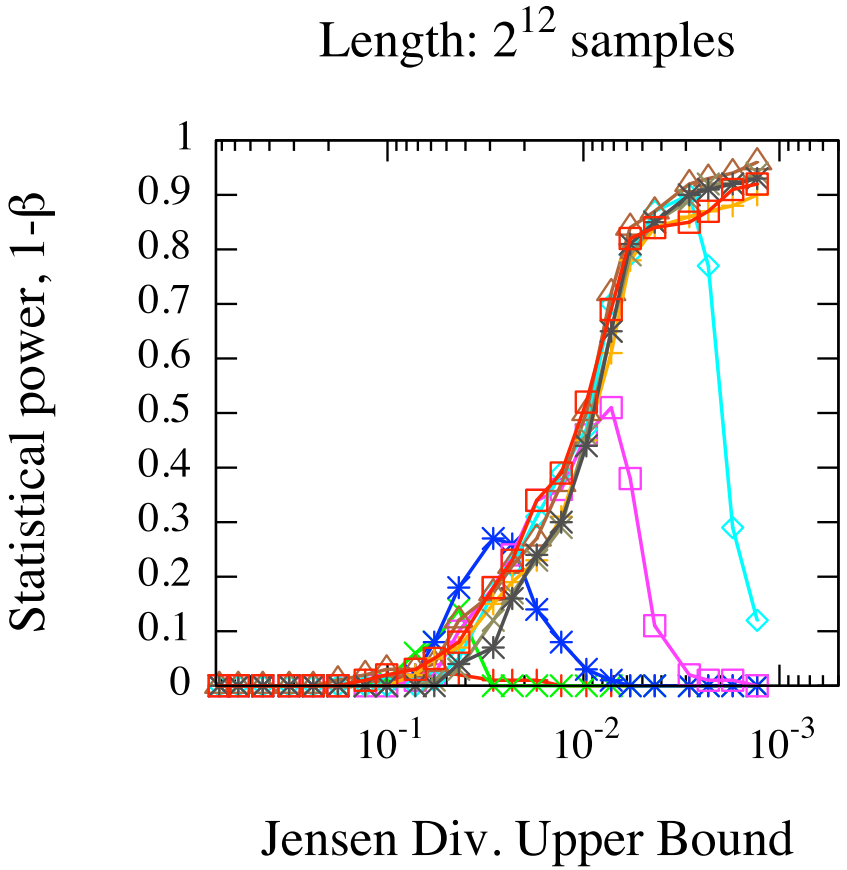


		Test accepts H_0	Test rejects H_0
Null hypothesis	H_0 **is true**	OK: $1-\alpha$ CL	α: **Type I Error**
Alternative hypothesis	H_1 **is true**	β: **Type II Error**	OK: $1-\beta$ Power

The goal is to maximize the power $1-\beta$

Statistical power: $\phi \leq 0.1$

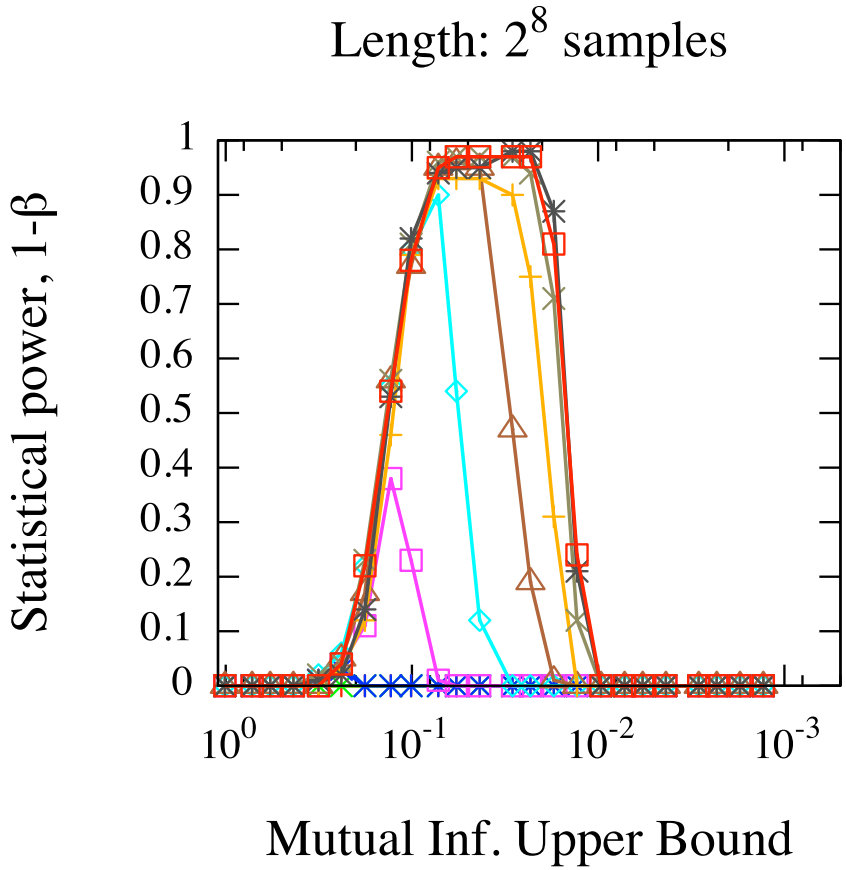


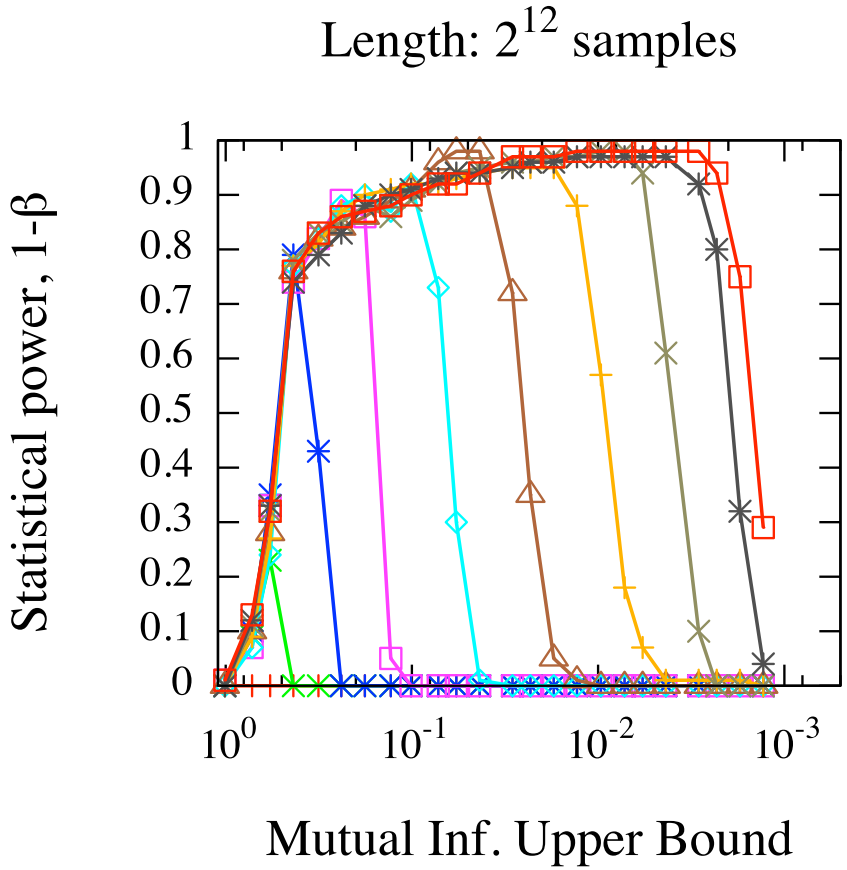


Preliminary results

Statistical power: $\phi \leq 0.1$

UNIVERSITY OF BIRMINGHAM





Preliminary results

$$\mathcal{I}^{\dagger}(\mathbf{X},\mathbf{Y}) = \frac{H(\mathbf{X}) - \mathcal{I}(\mathbf{X},\mathbf{Y})}{H(\mathbf{X})}$$

Mobility patterns: summary

UNIVERSITY OF BIRMINGHAM

- Mobility traces $\longrightarrow$ multivariate time series
- Information theoretical measures can be used to estimate the **similarity of different mobility patterns**
- **Jensen-Shannon divergence** and **Mutual Information**-based measures are suitable candidates
- **High statistical power** for suitable selection of upper/lower bounds
- The method is still valid for univariate time series and can be reliably adopted for other studies (finance, brain, genetics?)

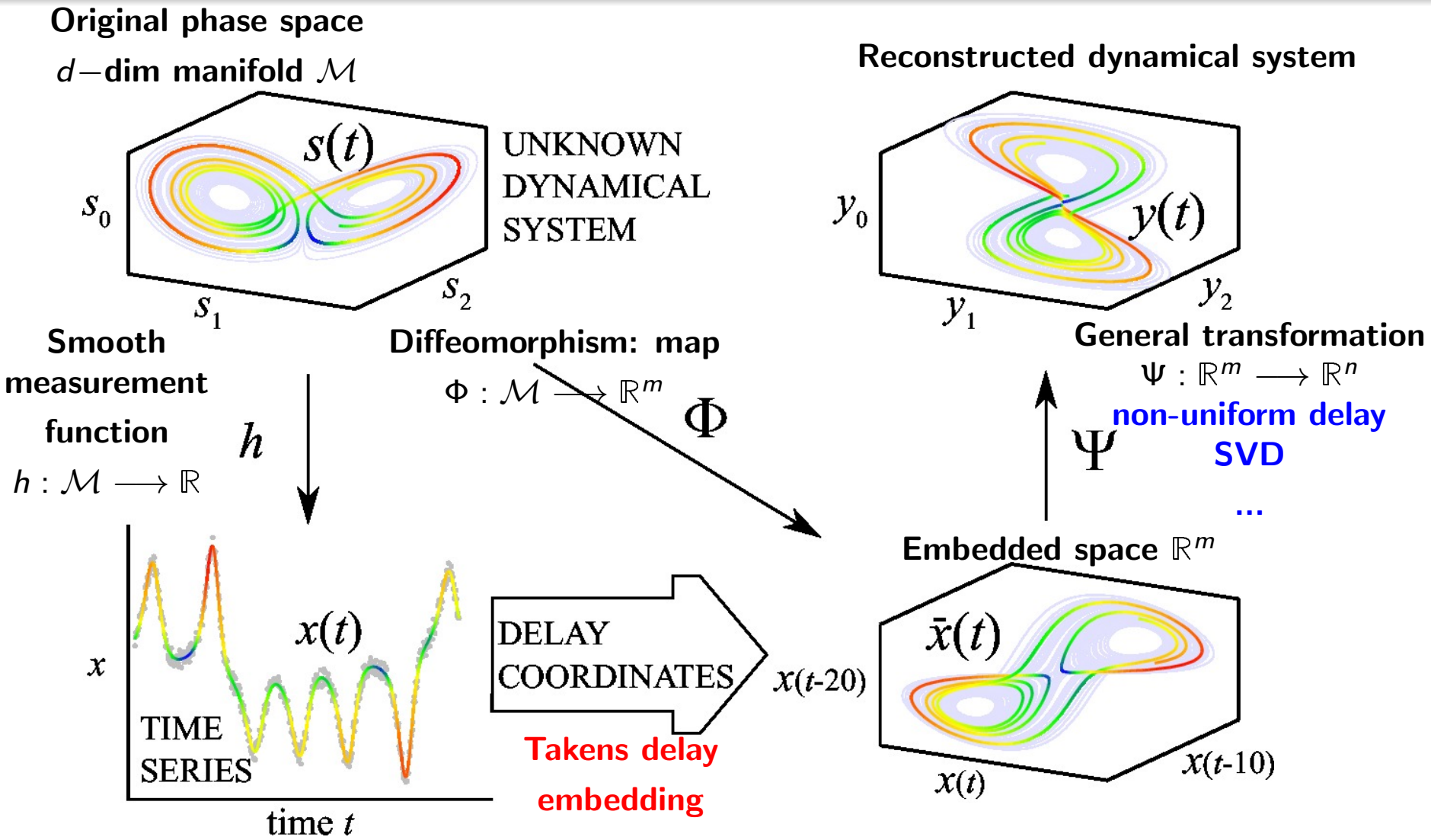
Idea: prediction of human mobility is important for several reasons. Under the assumptions of this study, we can try to **predict** movements by means of nonlinear methods from chaos theory. Could movements of users with similar mobility patterns be predicted with better accuracy?

Mobility patterns: summary

- Mobility traces $\longrightarrow$ multivariate time series
- Information theoretical measures can be used to estimate the **similarity of different mobility patterns**
- **Jensen-Shannon divergence** and **Mutual Information**-based measures are suitable candidates
- **High statistical power** for suitable selection of upper/lower bounds
- The method is still valid for univariate time series and can be reliably adopted for other studies (finance, brain, genetics?)

Idea: prediction of human mobility is important for several reasons. Under the assumptions of this study, we can try to **predict** movements by means of nonlinear methods from chaos theory. Could movements of users with similar mobility patterns be predicted with better accuracy?

Embeddology



* Picture readapted from L.C. Uzal et al, PRE (2011)

Delay Embedding: Reconstruction

- Univariate time series $x_n \longrightarrow m-$dimensional space preserving dynamical characteristics of the original phase space.
- Delay vector $\mathbf{x}_n$ from delayed measurements:

$$\mathbf{x}_n \equiv (x_{n-(m-1)\tau}, x_{n-(m-2)\tau}, \cdots, x_{n-\tau}, x_n)$$

- Reconstruction depends on two parameters m and τ (time delay) to be estimated
- Extension to multivariate observation $\mathbf{y}_n \equiv (y_{1,n}, y_{2,n}, \ldots, y_{M,n})$

$$\begin{aligned}\mathbf{v}_n \equiv\ & (y_{1,n-(m_1-1)\tau_1}, y_{1,n-(m_1-2)\tau_1}, \cdots, y_{1,n}, \\ & y_{2,n-(m_2-1)\tau_2}, y_{2,n-(m_2-2)\tau_2}, \cdots, y_{2,n}, \\ & \ldots \\ & y_{M,n-(m_M-1)\tau_M}, y_{1,n-(m_M-2)\tau_M}, \cdots, y_{M,n})\end{aligned}$$

- Reduce complexity: consider uniform embedding ($\tau_i = \tau$, $m_i = m$)

Delay Embedding: Reconstruction

- Univariate time series $x_n \longrightarrow m-$dimensional space preserving dynamical characteristics of the original phase space.
- Delay vector $\mathbf{x}_n$ from delayed measurements:

$$\mathbf{x}_n \equiv (x_{n-(m-1)\tau}, x_{n-(m-2)\tau}, \cdots, x_{n-\tau}, x_n)$$

- Reconstruction depends on two parameters m and τ (time delay) to be estimated

- Extension to multivariate observation $\mathbf{y}_n \equiv (y_{1,n}, y_{2,n}, ..., y_{M,n})$

$$\begin{aligned}\mathbf{v}_n \equiv\ & (y_{1,n-(m_1-1)\tau_1}, y_{1,n-(m_1-2)\tau_1}, \cdots, y_{1,n}, \\ & y_{2,n-(m_2-1)\tau_2}, y_{2,n-(m_2-2)\tau_2}, \cdots, y_{2,n}, \\ & \cdots \\ & y_{M,n-(m_M-1)\tau_M}, y_{1,n-(m_M-2)\tau_M}, \cdots, y_{M,n})\end{aligned}$$

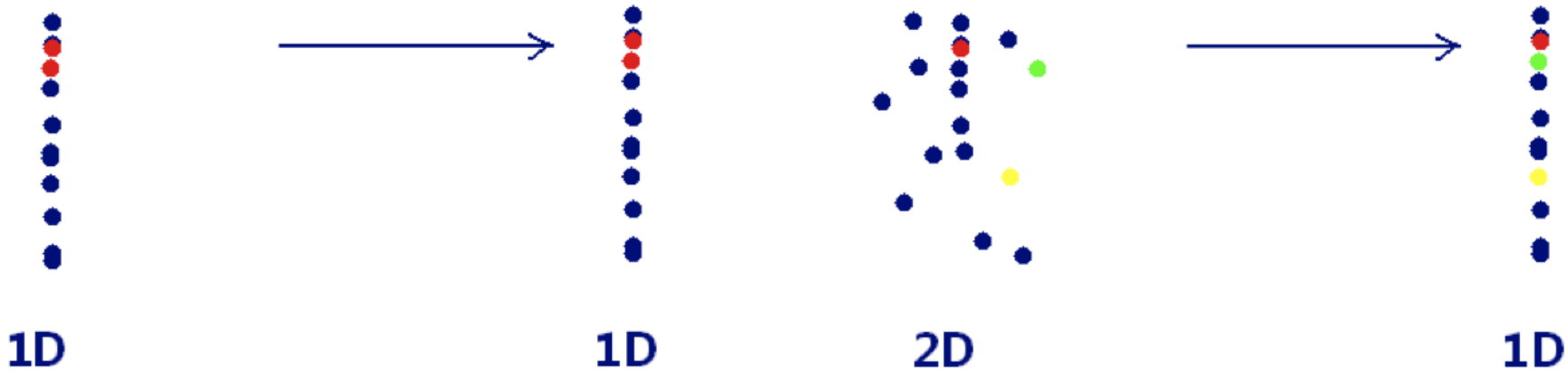
- Reduce complexity: consider uniform embedding ($\tau_i = \tau$, $m_i = m$)

The optimal delay $\tau_\star$ minimizes the self information of the time series (A. Fraser and H. Swinney, PRA (1986))

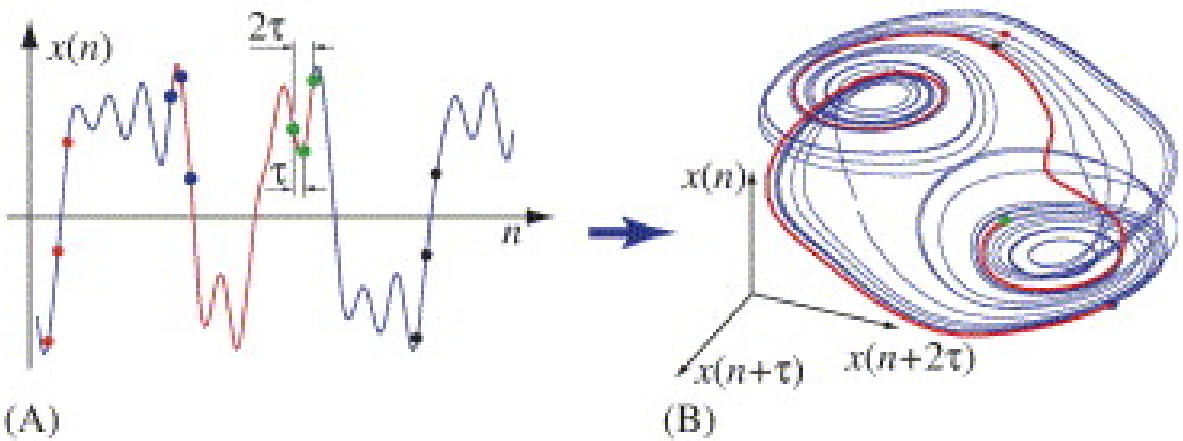
$\mathcal{I}(\tau)$ quantifies the amount of information about $x_{n+\tau}$ if x_n is known. In practice, choose $\tau_\star$ as the first local minimum of $\mathcal{I}(\tau)$

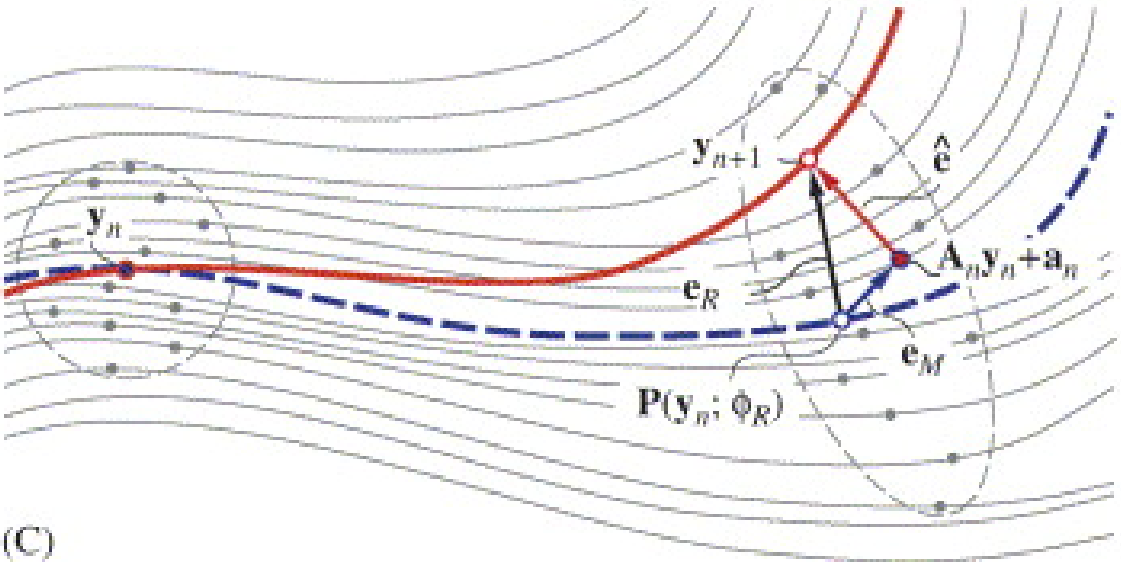
$$\mathcal{I}(\tau) = \sum_{ij} p_{ij}(\tau) \log \frac{p_{ij}(\tau)}{p_i(\tau) p_j(\tau)}$$

Build any embedding space from $m \geq 1$. The optimal embedding $m_\star$ minimizes the fraction of false nearest neighbors w.r.t. to $m_\star - 1$ (M. Kennel et al, PRA (1992); R. Hegger and H. Kantz, PRE (1999))



In practice, choose $m_\star$ if FNN($m_\star$) is smaller than a threshold, generally 5%: this guarantees that 95% of the phase space is well reconstructed





Approximate the dynamics locally in the phase space by a constant (M. Casdagli, Physica D (1989))

$\mathcal{U}_n$ is the neighbourhood of state $\mathbf{y}_n$ at time n

Forecast $\hat{\mathbf{y}}_{n+k}$ for $\mathbf{y}_{n+k}$:

$$\hat{\mathbf{y}}_{n+k} = \frac{1}{|\mathcal{U}_n|} \sum_{\mathbf{y}_j \in \mathcal{U}_n} \mathbf{y}_{j+k}$$

i.e., the average over the states which correspond to measurements k steps ahead of the neighbours $\mathbf{y}_j$

Picture from Dingwell et al, J. Biom. (2007)

Is it possible to predict human mobility?

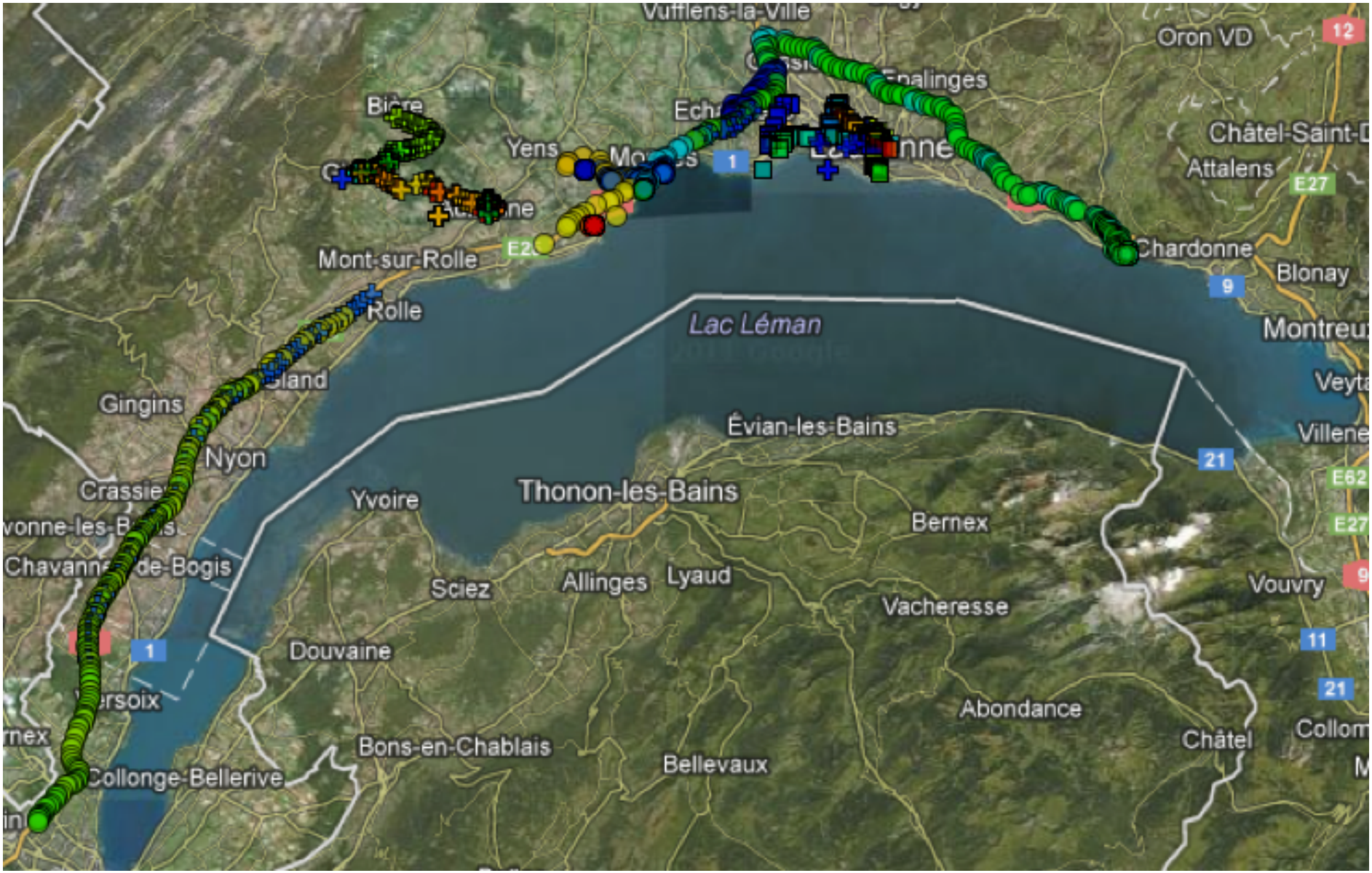
Nokia MDC dataset

- The complete dataset contains information from 152 smartphones (Nokia N95) for a year: address book, GPS, WLAN and Bluetooth traces, calls and SMS logs
- Individuals are students in Lausanne, Switzerland

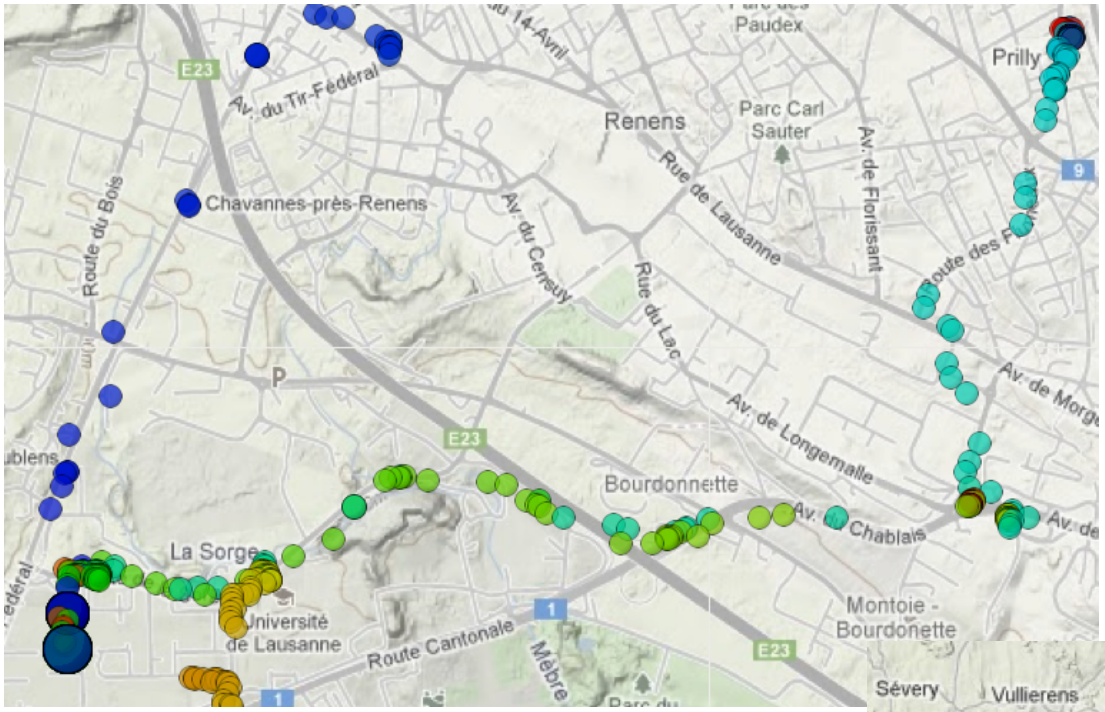
Joint work with A. Lima and M. Musolesi

- Our team received data from 39 devices, 14 phone numbers were missing. We analysed a subset of the data related to 25 devices
- We tried to predict the next place where an individual is moving to, by using his/her historical GPS readings

Interdependence and Predictability of Human Mobility and Social Interactions (winner)
M.D.D., A. Lima and M. Musolesi
Proc. of the Nokia MDC Workshop. Colocated with Pervasive 2012. Newcastle, UK. June 2012



Copyright for the map: 2012 TerraMetrics, Map data 2012 Google, Tele Atlas



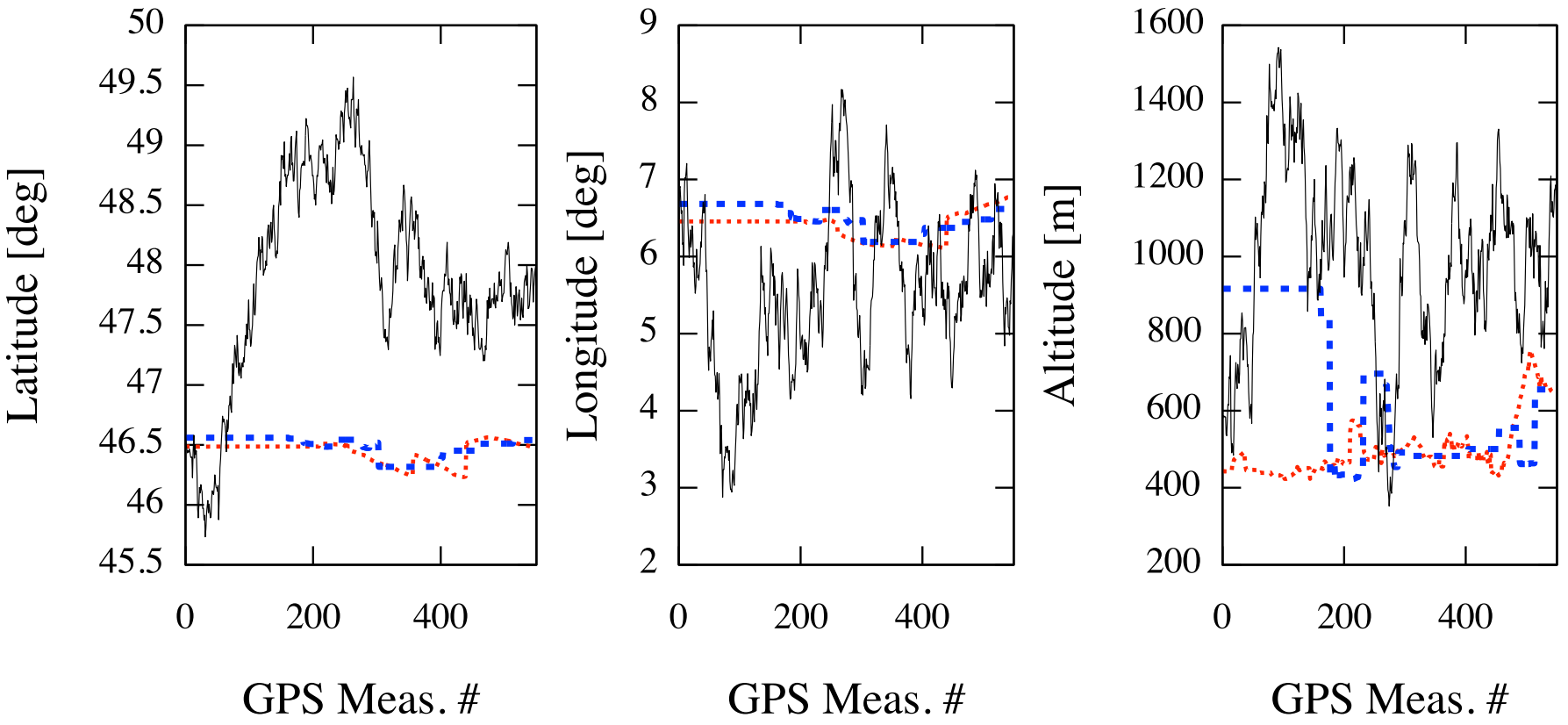
- GPS readings are not evenly sampled $\longrightarrow$ problem for embedding reconstruction
- **Our nonlinear mobility model:**

$$\dot{\mathbf{x}}(t) = \underbrace{\mathbf{f}[\mathbf{x}(t), t]}_{\text{Nonlinear dynamics}} + \underbrace{\eta(t)}_{\text{Noise term}}$$

where the multivariate time series is given by

$$\mathbf{x}(t) \equiv \begin{pmatrix} h(t) \\ \phi(t) \\ \lambda(t) \\ \xi(t) \end{pmatrix} \begin{matrix} \text{Hour of the day} \\ \text{Latitude} \\ \text{Longitude} \\ \text{Altitude} \end{matrix}$$

NMDC: One-user multivariate prediction



Dotted/Red: Observation for user 179
Solid/Black: linear prediction by multivariate ARMA
Dashed/Blue: prediction by multivariate nonlinear predictor

$\Longrightarrow$ Cumulative rms error $\approx 0.2^\circ$ on lat/long

Nodes	Social Link	Position: Cum. rms error [deg]	Altitude: Cum. rms error [m]
026, 127	None	0.167	66.33
063, 123	Present	0.011	20.95
094, 009	Present	0.003	5.57

- Prediction error for nodes with no social contacts $\approx$ as one-user prediction
- If social ties are present, prediction considerably improves

Intriguing result but **NOT** definitive:

- **Lack of statistics**
- Possible **biased** dataset (individuals are all students, etc)

Is it possible to predict human mobility?

Application to **my** mobility pattern...

Input

$\Longrightarrow$

Prediction from the algorithm

The place where I was for the last 10 years, in July (my home in Messina, Sicily)

Predicted

VS

Observed

The place where I was for the last 10 years, in July (my home in Messina, Sicily)

The place where I am **this** July

Still a lot of work to do! :-)

Questions?

Manlio De Domenico

E-Mail: m.dedomenico@cs.bham.ac.uk

Homepage: http://www.cs.bham.ac.uk/~dedomenm/