Interdependence and predictability of human mobility and social interactions

M. De Domenico

\textsuperscript{1}School of Computer Science, University of Birmingham, UK

Mathematics of Networks
University of Warwick, 20\textsuperscript{th} July 2012
My background...

Astroparticle physics
Cosmic rays

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

I come from Italy...

In UK since March...

Complex systems
Network science
Nonlinear time series analysis
Outline of this talk

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 Challenge
Interdependence

(Unknown) network of dynamical systems

Real-world

Observations! ⇦
Network from time series: financial market

**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))

M. De Domenico Interdependence and predictability of human mobility and social interactions

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

D. Garlaschelli et al, Physica A (2005)
**Goal:** investigate brain connectivity to detect unhealthy (epileptic) subjects

**Time series**

- MEG signals
- M. Chavez et al., PRL (2010)

**Phase-Locking Values**

- as measure of correlation/synchronization

**PLV_{j,k}(t_0, f_0)**

- node j
- node k

M. Chavez et al., PRL (2010)
What about human movements?

**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?

- 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 \(\rightarrow\) 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?

- 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 $\rightarrow$ 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
Let $X$ be a multivariate stochastic variable (e.g., GPS reading)

$P_X(x)$ is the true probability density function (PDF)

$Q_X(x)$ is some approximate model describing outcomes of $X$

**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)

$$D_{KL}(P||Q) = \sum_{x \in X} P_X(x) \log \frac{P_X(x)}{Q_X(x)} = \mathcal{H}(P, Q) - \mathcal{H}(P)$$

- $D_{KL}(P||Q) \geq 0$ (not bounded above: undesirable feature)
- $D_{KL}(P||Q) = 0 \iff P_X(x) = Q_X(x)$
- $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ (undesirable feature)
Let $X$ be a multivariate stochastic variable (e.g., GPS reading)

$P_X(x)$ is the true probability density function (PDF)

$Q_X(x)$ is some approximate model describing outcomes of $X$

**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)

$$D_{KL}(P||Q) = \sum_{x \in X} P_X(x) \log \frac{P_X(x)}{Q_X(x)} = H(P, Q) - H(P)$$

- $D_{KL}(P||Q) \geq 0$ (not bounded above: *undesirable feature*)
- $D_{KL}(P||Q) = 0 \iff P_X(x) = Q_X(x)$
- $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ (*undesirable feature*)
Information theoretical measures

Possible alternative is the Jensen-Shannon divergence

\[ D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||R) + \frac{1}{2} D_{KL}(Q||R), \quad R = \frac{P + Q}{2} \]

- **Bounded:** \( 0 \leq D_{JS}(P||Q) \leq 1 \)
- \( D_{JS}(P||Q) = 0 \Leftrightarrow P_X(x) = Q_X(x) \)
- **Symmetric:** \( D_{JS}(P||Q) = D_{JS}(Q||P) \)

Application to human mobility

- \( X \) represents the motion of a user on the Earth
- Random samples \( x \) (\( y \)) drawn from \( X \) (\( Y \)) correspond to geographic coordinates
- The PDF of \( x \) (\( y \)) quantifies the fraction of time spent by the user \( X \) (\( Y \)) in a particular position
- Use \( D_{JS}(P_X||P_Y) \) and \( D_{KL}(P_X||P_Y) \) to quantify the similarity of their mobility patterns
Mutual information is another way of measuring the correlation:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)} = D_{KL}(P_{XY} \parallel P_XP_Y)$$

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

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

- 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 $\Rightarrow$ 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: $k = 1, 2, \ldots, N$
Statistical validation of the method

- 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 $\Rightarrow$ 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: $k = 1, 2, \ldots, N$
Statistical validation of the method

Procedure

1. Consider 20 chaotic agents, fixing number of samples and SNR
2. Generate $N$ random undir. unw. $20 \times 20$ 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 $0 < \phi < 1$. 

M. De Domenico Interdependence and predictability of human mobility and social interactions
Statistical validation of the method

**Procedure**

1. Consider 20 chaotic agents, fixing number of samples and SNR
2. Generate $N$ random undir. unw. $20 \times 20$ 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 $0 < \phi \leq 1$. 

M. De Domenico Interdependence and predictability of human mobility and social interactions
**Correlation matrices**

\[
\begin{align*}
(H_A - I_{A,B})/H_A & \\
(H_A - H_B)/H_A &
\end{align*}
\]

- **NodeID vs NodeID**

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

\[
b_{ij} = \begin{cases} 
1 & \text{if } b_{ij}^* \geq (\leq) \text{ thresh} \\
0 & \text{if } b_{ij}^* < (>) \text{ thresh}
\end{cases}
\]
Our numerical studies show Jensen-Shannon divergence and Mutual Information are the most promising.
Hypothesis Testing: Statistical Errors

- Accept $H_0$
- Reject $H_0$

**Null hypothesis**: $H_0$ is true
**Alternative hypothesis**: $H_1$ is true

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

The goal is to **maximize** the power $1 - \beta$
Statistical power: \( \phi \leq 0.1 \)

**Preliminary results**

Length: \( 2^8 \) samples

Length: \( 2^{12} \) samples

Jensen Div. Upper Bound

Jensen Div. Upper Bound
Statistical power: $\phi \leq 0.1$

Preliminary results

\[ I^\dagger(X, Y) = \frac{H(X) - I(X, Y)}{H(X)} \]
Mobility traces $\rightarrow$ 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 → 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**

Original phase space
\[ d-\text{dim manifold } \mathcal{M} \]

- **s(t)**
- **s_0**
- **s_1**
- **s_2**

**UNKNOWN DYNAMICAL SYSTEM**

**Smooth measurement function**
\[ h : \mathcal{M} \rightarrow \mathbb{R} \]

**Diffeomorphism: map**
\[ \Phi : \mathcal{M} \rightarrow \mathbb{R}^m \]

**Reconstructed dynamical system**

- **y_0**
- **y_1**
- **y_2**

**General transformation**
\[ \Psi : \mathbb{R}^m \rightarrow \mathbb{R}^n \]
- **non-uniform delay**
- **SVD**
- **...**

**Embedded space** \[ \mathbb{R}^m \]

**TIME SERIES**

- **x(t)**
- **x**

**DELAY COORDINATES**

- **x(t-20)**
- **\tilde{x}(t)**

**Takens delay embedding**

* Picture readapted from L.C. Uzal et al, PRE (2011)
Delay Embedding: Reconstruction

- Univariate time series $x_n \rightarrow m$-dimensional space preserving dynamical characteristics of the original phase space.

- Delay vector $x_n$ from delayed measurements:

$$x_n \equiv (x_n-(m-1)\tau, x_n-(m-2)\tau, \ldots, x_n-\tau, x_n)$$

- Reconstruction depends on two parameters $m$ and $\tau$ (time delay) to be estimated

- Extension to multivariate observation $y_n \equiv (y_{1,n}, y_{2,n}, \ldots, y_{M,n})$

$$v_n \equiv (y_{1,n}-(m_1-1)\tau_1, y_{1,n}-(m_1-2)\tau_1, \ldots, y_{1,n},$$

$$y_{2,n}-(m_2-1)\tau_2, y_{2,n}-(m_2-2)\tau_2, \ldots, y_{2,n},$$

$$\ldots$$

$$y_{M,n}-(m_M-1)\tau_M, y_{1,n}-(m_M-2)\tau_M, \ldots, y_{M,n})$$

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

- Univariate time series $x_n \rightarrow m$-dimensional space preserving dynamical characteristics of the original phase space.

- Delay vector $x_n$ from delayed measurements:
  $$x_n \equiv (x_{n-(m-1)\tau}, x_{n-(m-2)\tau}, \ldots, x_{n-\tau}, x_n)$$

- Reconstruction depends on two parameters $m$ and $\tau$ (time delay) to be estimated.

- Extension to multivariate observation $y_n \equiv (y_{1,n}, y_{2,n}, \ldots, y_{M,n})$
  $$v_n \equiv (y_{1,n-(m_1-1)\tau_1}, y_{1,n-(m_1-2)\tau_1}, \ldots, y_{1,n},$$
  $$y_{2,n-(m_2-1)\tau_2}, y_{2,n-(m_2-2)\tau_2}, \ldots, y_{2,n},$$
  $$\ldots$$
  $$y_{M,n-(m_M-1)\tau_M}, y_{1,n-(m_M-2)\tau_M}, \ldots, y_{M,n})$$

- 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))

$I(\tau)$ 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 $I(\tau)$

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

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))

\[ U_n \text{ is the neighbourhood of state } y_n \text{ at time } n \]

Forecast \( \hat{y}_{n+k} \) for \( y_{n+k} \):

\[ \hat{y}_{n+k} = \frac{1}{|U_n|} \sum_{y_j \in U_n} y_{j+k} \]

i.e., the average over the states which correspond to measurements \( k \) steps ahead of the neighbours \( y_j \)

Picture from Dingwell et al, J. Biom. (2007)
The Nokia Mobile Data Challenge

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 maps: 2012 TerraMetrics, Map data 2012 Google, Tele Atlas
GPS readings are not evenly sampled \(\rightarrow\) problem for embedding reconstruction

Our nonlinear mobility model:

\[
\dot{x}(t) = f[x(t), t] + \eta(t)
\]

where the multivariate time series is given by

\[
x(t) \equiv \begin{pmatrix}
h(t) \\
\phi(t) \\
\lambda(t) \\
\xi(t)
\end{pmatrix}
\]

- Hour of the day
- Latitude
- Longitude
- Altitude
Dotted/Red: Observation for user 179
Solid/Black: linear prediction by multivariate ARMA
Dashed/Blue: prediction by multivariate nonlinear predictor

⇒ Cumulative rms error ≈ 0.2° on lat/long
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...
The place where I was for the last 10 years, in July (my home in Messina, Sicily)
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/

My own pictures are released under CC BY 3.0: http://creativecommons.org/licenses/by/3.0/
Give credits to: M. De Domenico, University of Birmingham