

# Nonlinear Characterization of Activity Dynamics in Online Collaboration Websites

Tiago Santos<sup>1</sup>   Simon Walk<sup>2</sup>   Denis Helic<sup>3</sup>

<sup>1</sup>Know-Center, Graz, Austria

<sup>2</sup>Stanford University

<sup>3</sup>Graz University of Technology

3. April 2017

# Motivation

Success of online collaboration websites depends critically on content contributed by users.

- For example, StackExchange vs. Google knol.

**Problem:** Key deciding factors of success and failure of online collaboration websites?

**Goal:** Uncover hidden nonlinear behavior in activity dynamics.

# Nonlinear Time Series Analysis and its Applications

Nonlinear time series analysis studies reconstructions of high dimensional dynamical systems from low dimensional ones.

Example applications:

- Small and Tse [1] predicted the outcome of a roulette wheel.
- Hsieh [2] found nonlinear behavior in stock returns.
- Strozzi et al. [3] detected events in the stock market.
- More examples in Marwan et al. [4] and Bradley and Kantz [5].

New application: activity dynamics in online collaboration websites.

# Dynamical Systems for Networks

Dynamical systems provide mathematical formalizations for the evolution of numerical quantities over time [6, 7].

Applications of dynamical system theory to study activity in networks:

- Ribeiro [8] models daily active users in online communities with behavior of active and inactive users.
- Walk et al. [9] model activity in collaboration networks with activity decay rate and peer influence growth.

Our approach: Characterize activity by its propensity to have originated in a dynamical system

# Nonlinearity Tests

We assess nonlinearity with 9 statistical tests:

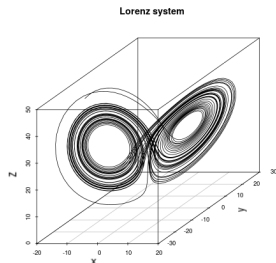
- AR process-based tests:
  - Broock, Dechert and Scheinkman [BDS] test [10] on ARIMA residuals
  - Keenan's one-degree test for nonlinearity [11]
  - McLeod-Li test [12]
  - Tsay's test for nonlinearity [13]
  - Likelihood ratio test for threshold nonlinearity [14]
- Neural Networks-based tests:
  - Teraesvirta's neural network test [15]
  - White neural network test [16]
- Other tests:
  - Wald-Wolfowitz runs test [17] on the number of times time series grows
  - Surrogate test - time asymmetry [18]

# Reconstructing nonlinear dynamical system from univariate time series

We reconstruct state space with Takens' embedding theorem [19] to get:

$$R_t = (x_t, x_{t-\tau}, x_{t-2\tau}, \dots, x_{t-(m-1)\tau}) \in \mathbb{R}^m. \quad (1)$$

Example: Lorenz dynamical system

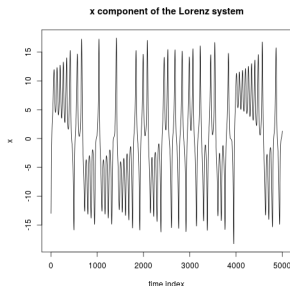
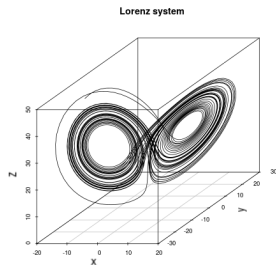


# Reconstructing nonlinear dynamical system from univariate time series

We reconstruct state space with Takens' embedding theorem [19] to get:

$$R_t = (x_t, x_{t-\tau}, x_{t-2\tau}, \dots, x_{t-(m-1)\tau}) \in \mathbb{R}^m. \quad (1)$$

Example: Lorenz dynamical system



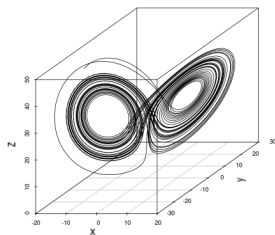
# Reconstructing nonlinear dynamical system from univariate time series

We reconstruct state space with Takens' embedding theorem [19] to get:

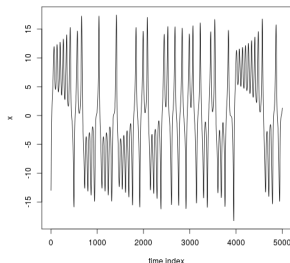
$$R_t = (x_t, x_{t-\tau}, x_{t-2\tau}, \dots, x_{t-(m-1)\tau}) \in \mathbb{R}^m. \quad (1)$$

Example: Lorenz dynamical system

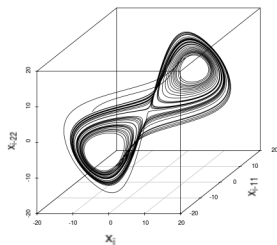
Lorenz system



x component of the Lorenz system



First 3 dimensions of Lorenz's reconstructed phase space





# Forecasting univariate time series

We employ linear and nonlinear models to forecast time series.

- Linear models:
  - Linear regression: linear combination of Fourier terms and trend
  - ARIMA models: differenced, linear combination of auto-regressors and lagged moving average error terms
  - Exponential Smoothing (ETS) models: linear combination of lagged terms, such as level, trend, seasonality and error

# Forecasting univariate time series

We employ linear and nonlinear models to forecast time series.

- Linear models:
  - Linear regression: linear combination of Fourier terms and trend
  - ARIMA models: differenced, linear combination of auto-regressors and lagged moving average error terms
  - Exponential Smoothing (ETS) models: linear combination of lagged terms, such as level, trend, seasonality and error
- Nonlinear models:
  - Reconstruct dynamical system properties with Takens embedding
  - Forecast univariate time series by following nearby trajectories in reconstructed state space

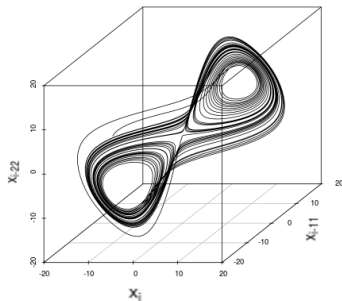
# Recurrence Analysis

Analyze reconstructed state spaces with Recurrence Plots:

$$R_{i,j}(\epsilon) = \Theta(\epsilon - \|\vec{x}_i - \vec{x}_j\|). \quad (2)$$

Example: Lorenz dynamical system

First 3 dimensions of Lorenz's reconstructed phase space



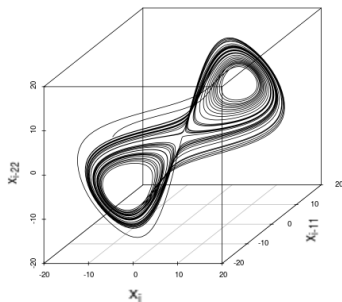
# Recurrence Analysis

Analyze reconstructed state spaces with Recurrence Plots:

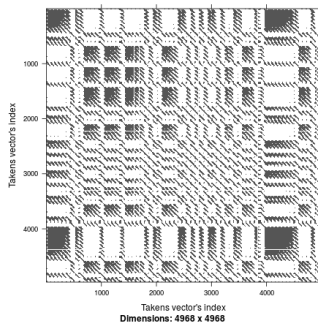
$$R_{i,j}(\epsilon) = \Theta(\epsilon - \|\vec{x}_i - \vec{x}_j\|). \quad (2)$$

Example: Lorenz dynamical system

First 3 dimensions of Lorenz's reconstructed phase space



Lorenz recurrence plot



# Datasets, pre-processing & experiment configuration

Datasets and pre-processing:

- Datasets: 16 randomly picked StackExchange Q&A portals
- Activity-based time series: number of questions, answers and comments per user per day
- Pre-processing: weekly aggregation, burn-in, detrend

# Datasets, pre-processing & experiment configuration

## Datasets and pre-processing:

- Datasets: 16 randomly picked StackExchange Q&A portals
- Activity-based time series: number of questions, answers and comments per user per day
- Pre-processing: weekly aggregation, burn-in, detrend

## Test and forecast setup:

- Significance level of the 9 tests for nonlinearity: 95%
- Categorize datasets on number of tests indicating nonlinearity
- Forecast 1 year of activity for all datasets
- Compare forecast root mean squared error (RMSE) of the 4 models

# Nonlinearity test results and forecast performance comparison

Group datasets on nonlinearity test results and rank forecast RMSE per group with the Friedman test [20]:

- 10 out of 16 datasets with  $\leq 4/9$  tests indicating nonlinearity.  
Friedman test rank: 1. ETS, 2. ARIMA, 3. Nonlinear, 4. Linear
- 6 out of 16 datasets with  $\geq 5/9$  tests indicating nonlinearity.  
Friedman test rank: 1. Nonlinear, 2. ARIMA, 2. ETS, 4. Linear

# Nonlinearity test results and forecast performance comparison

Group datasets on nonlinearity test results and rank forecast RMSE per group with the Friedman test [20]:

- 10 out of 16 datasets with  $\leq 4/9$  tests indicating nonlinearity.  
Friedman test rank: 1. ETS, 2. ARIMA, 3. Nonlinear, 4. Linear
- 6 out of 16 datasets with  $\geq 5/9$  tests indicating nonlinearity.  
Friedman test rank: 1. Nonlinear, 2. ARIMA, 2. ETS, 4. Linear

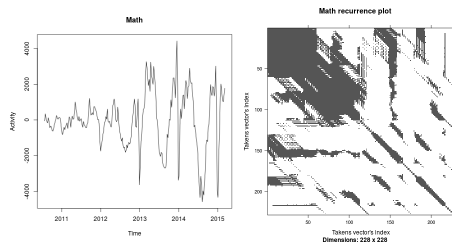
Observations:

- Neural network-based tests are more sensitive to nonlinear dynamics
- Presence of nonlinear dynamics impacts forecast and modeling efforts



# Recurrence Plot Analysis

Study reconstructed state spaces of 2 datasets deemed nonlinear:

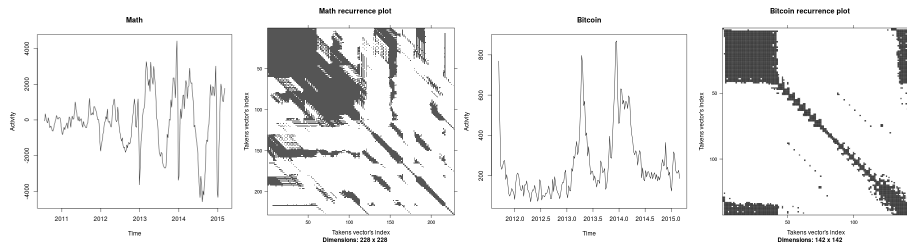


The RP empowers activity dynamics modeling efforts:

- Math: Drift, chaotic dynamics and slowly changing states

# Recurrence Plot Analysis

Study reconstructed state spaces of 2 datasets deemed nonlinear:



The RP empowers activity dynamics modeling efforts:

- Math: Drift, chaotic dynamics and slowly changing states
- Bitcoin: Periodic dynamics and non-stationary transitions

# Conclusions & Future Work

## Conclusions:

- Group activity-based time series by propensity to originate from dynamical systems
- Increase accuracy in activity forecast experiments
- Customize activity models with Recurrence Plots
- More and longer time series → more conclusive results

# Conclusions & Future Work

## Conclusions:

- Group activity-based time series by propensity to originate from dynamical systems
- Increase accuracy in activity forecast experiments
- Customize activity models with Recurrence Plots
- More and longer time series → more conclusive results

## Future work:

- Understand *reason* for differences in nonlinear behavior
- Study underlying collaboration networks
- Recurrence Quantification Analysis





# Questions?

Thank you very much for your time!  
Questions?






## Results table

Dataset	Weeks	$\tau$	$m$	Nonlin. test score	Positive nonlin. tests	ARIMA	ETS	Linear	Nonlin.
english <sup>b</sup>	240	2	9	2/9	[2] [13]	0.6794	0.4452	0.3329	0.3080
unix <sup>b</sup>	239	1	7	2/9	[2] [13]	0.2091	0.2092	0.2418	0.2074
chemistry <sup>b</sup>	158	2	7	3/9	[2] [13] [4]	0.4982	0.2539	0.3247	0.4610
webmasters	244	1	8	3/9	[9] [13] [15]	0.2313	0.2528	0.3341	0.2346
chess	148	2	8	4/9	[2] [9] [13] [15]	0.2545	- <sup>a</sup>	0.5622	0.5110
history	177	1	9	4/9	[2] [9] [13] [4]	0.3503	0.2368	0.3044	0.4052
linguistics	181	2	6	4/9	[2] [9] [13] [15]	0.2512	0.2704	0.3009	0.3280
sqa	200	3	9	4/9	[2] [9] [13] [15]	1.8136	0.2531	0.6549	0.3903
tex <sup>b</sup>	241	1	7	4/9	[13] [21] [4] [15]	0.1589	0.1580	0.2767	0.2751
tridion	107	1	7	4/9	[19] [10] [9] [13]	0.2717	- <sup>a</sup>	0.6144	- <sup>a</sup>
Friedman test rank of models' forecast RMSE on datasets with nonlin. test score < 5/9						2	1	4	3
arduino	56	1	10	5/9	[2] [19] [10] [9] [13]	0.3489	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>
sports	159	1	7	5/9	[2] [9] [13] [4] [15]	0.2442	0.3348	0.4019	0.3323
ux	239	2	8	5/9	[2] [10] [9] [13] [21]	0.3479	0.1743	0.3491	0.1374
bitcoin	182	4	11	6/9	[2] [19] [10] [9] [13] [15]	0.6099	0.5549	0.5938	0.5781
math <sup>b</sup>	242	2	8	6/9	[2] [19] [13] [21] [4] [15]	0.1327	0.2314	0.3521	0.2912
bicycles	235	2	7	7/9	[2] [19] [10] [9] [13] [4] [15]	0.2971	0.3097	0.3252	0.2805
Friedman test rank of models' forecast RMSE on datasets with nonlin. test score $\geq$ 5/9						2 <sup>c</sup>	2 <sup>c</sup>	4	1

# References I






-  Small, M and Tse, CK. Predicting the outcome of roulette. *Chaos: an interdisciplinary journal of nonlinear science* 2012;22:033150.
-  Hsieh, DA. Chaos and nonlinear dynamics: application to financial markets. *The journal of finance* 1991;46:1839–1877.
-  Strozzi, F, Zaldívar, JM, and Zbilut, JP. Application of nonlinear time series analysis techniques to high-frequency currency exchange data. *Physica A: Statistical Mechanics and its Applications* 2002;312:520–538.
-  Marwan, N, Romano, MC, Thiel, M, and Kurths, J. Recurrence plots for the analysis of complex systems. *Physics reports* 2007;438:237–329.

## References II






-  Bradley, E and Kantz, H. Nonlinear time-series analysis revisited. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 2015;25:097610.
-  Luenberger, DGDG. Introduction to dynamic systems; theory, models, and applications. *Tech. rep.* 1979.
-  Guckenheimer, J and Holmes, PJ. Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. *Vol. 42. Springer Science & Business Media*, 2013.
-  Ribeiro, B. Modeling and predicting the growth and death of membership-based websites. *In: Proceedings of the 23rd international conference on World Wide Web. ACM.* 2014:653–664.
-  Walk, S, Helic, D, Geigl, F, and Strohmaier, M. Activity dynamics in collaboration networks. *ACM Transactions on the Web (TWEB)* 2016;10:11.



## References III

-  Broock, W, Scheinkman, JA, Dechert, WD, and LeBaron, B. A test for independence based on the correlation dimension. *Econometric reviews* 1996;15:197–235.
-  Keenan, DM. A Tukey nonadditivity-type test for time series nonlinearity. *Biometrika* 1985;72:39–44.
-  McLeod, AI and Li, WK. Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* 1983;4:269–273.
-  Tsay, RS. Nonlinearity tests for time series. *Biometrika* 1986;73:461–466.
-  Chan, KS. Percentage points of likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)* 1991:691–696.

# References IV

- 
 Teräsvirta, T, Lin, CF, and Granger, CW. Power of the neural network linearity test. *Journal of Time Series Analysis* 1993;14:209–220.
- 
 Lee, TH, White, H, and Granger, CW. Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics* 1993;56:269–290.
- 
 Siegel, S. Nonparametric statistics for the behavioral sciences. 1956.
- 
 Schreiber, T and Schmitz, A. Surrogate time series. *PhysicaD:NonlinearPhenomena* 2000;142:346–382.
- 
 Takens, F. Detecting strange attractors in turbulence. In: *Dynamical systems and turbulence, Warwick 1980*. Springer, 1981:366–381.

# References V



Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* 2006;7:1–30.