COMPUTATIONAL SOCIAL SCIENCE

Péter Érdi
perdi@kzoo.edu

Henry R. Luce Professor
Center for Complex Systems Studies
Kalamazoo College
http://people.kzoo.edu/ perdi/

and

Institute for Particle and Nuclear Physics, Wigner Research Centre, Hungarian Academy of Sciences, Budapest
http://cneuro.rmki.kfki.hu/
1. Towards a computational social science: from data mining via social simulations to prediction

2. Sociodynamics: from concepts to data and back

3. Prediction of emerging technologies based on analysis of the U.S patent citation network

4. Towards a Predictive Theory of the US National Budget

5. Conclusions
Towards a computational social science: from data mining via social simulations to prediction

Computational social science is the interdisciplinary studies of complex social systems by combining

- social theories
- data mining
- dynamical modeling techniques

Towards a computational social science: from data mining via social simulations to prediction

- to predict, control and management of socio-economic crises and political insabilities
- to avoid "pathological" collective behavior: panic, extremism, breakdown of trust …
- to understand and manage demographic changes
- to control of spreading epidemics
- to ensure security and peace
- to understand the structure and dynamics of different social networks
- to defend intellectual property rights
- to understand the relationship between individual and institutional decision making mechanism
Towards a computational social science: from data mining via social simulations to prediction

Figure 1: Instabilities, collective phenomena, social networks

Political instability index. Vulnerability to social and political unrest. From Economist Intelligence Unit

Panic: stadium collapses in Abidjan

Patent citation network: Chemistry: molecular biology and microbiology: from Steve Borgatti
Towards a computational social science: from data mining via social simulations to prediction

Figure 2: Visioneer is a European project aiming to reach a better, quantitative understanding of complex socio-economic systems: Dirk Helbing: ETH, Zurich
Towards a computational social science: from data mining via social simulations to prediction

NEW SOURCES of DATA & (OLD) NEW methods

Figure 3: The quantity of information in the world is soaring. New methods and more educated people are necessary to extract/process information.

- internet
- sensor networks
- government databases
The world of Tycho Brahe: **DATA COLLECTION** (induction starts here)
Data, Rules, Prediction: Lessons from Tycho de Brahe, Kepler and Newton

Kepler: MATHEMATICAL not locally but globally predictive: integral laws.

Newton’s laws: PREDICTIVE gravitation + differential laws.
When an external force acts on a body of constant mass then the acceleration produced is directly proportional to the force.

The discovery of Neptun: Adams vs Le Verrie. Galle
GROWTH PROCESSES

Figure 4: Unbounded growth processes: infinite and finite-time singularities
From finite-time singularity to crash

- unbalanced (higher-than-linear) positive feedback \(\rightarrow\) finite-time singularities
- infinite value during finite time: (chemical) explosion and "explosion"
- earthquakes, volcano, epilepsy, stock prices ...
- super-exponential increase (due to the irrational expectations) cannot be continued for “ever” due to the unstable nature of this process
- followed by a compensatory process (i.e. stock market crash)

Figure 5: Stock market crashes
Interpreting Temporal Data: Growth, Saturation, Boom-and-Bust, Oscillation

**Basic Dynamics**

- **Linear Growth**
  - $v = \dot{x} = c$
  - External!
  - Stable: $a > 0$, unstable: $a < 0 \rightarrow$ extinction

- **Exponential Growth**
  - $v = ax$
  - Positive feedback
  - Unstable: $a > 0$

- **Super-Exponential Growth**
  - $v = ax^n$
  - $a > 0$, $n > 1$

- **Boundless Growth**

- **Pull-back Dynamic**
  - Equilibrium

- **Boom-and-Bust**
  - $\dot{x} = ax - bx^2$
  - Renewable cycle

- **Logistic Growth**
  - $\dot{x} = c - dx$
  - Non-renewable

- **Y: Resources**
  - $\dot{Y} = ax$
  - Non-renewable
  - $\dot{Y} = -ax + by$
  - Renewable
A) war dynamics

\[
\frac{dR}{dt} = -k_BRB + \alpha R \left(1 - \frac{R}{K}\right) \tag{1}
\]

\[
\frac{dB}{dt} = -k_RBR + \beta B \left(1 - \frac{B}{L}\right), \tag{2}
\]

where \(\alpha, \beta, K\) and \(L\) are positive constants.

Generalized Lotka-Volterra equation: a famous model of competition of the struggle for existence

Georgyi Frantsevitch Gause (1910-1986)
Temporal dynamics: interacting variables: equilibrium, oscillation, chaos

A) war dynamics

principle of competitive exclusion vs. “permanent war”,
Gause, Strogatz, Epstein

Figure 6: Phase planes of the possible outcomes of bilateral wars
B) drug propagation and control

(adapted from Behrens et. al, 2004)

Figure 7: Simple dynamic models of drug propagation
B) drug propagation and control

(adapted from Behrens et. al, 2004)

\[
A_{t+1} = (1 - \alpha)A_t + af(A_t, D_t), \quad A_0 = A(0), \quad \text{(3a)}
\]
\[
D_{t+1} = (1 - \beta)D_t + bf(A_t, D_t), \quad D_0 = D(0), \quad \text{(3b)}
\]

\(a, b > 0\), and \(\alpha, \beta \in [0, 1]\). The positive and negative feedback effects should be expressed in the form of the function \(f\).

A rather general form for the function is

\[
f(A, D) = \frac{1}{1 + e^{-c(A-D)}}, \quad c > 0. \quad \text{(4)}
\]
Temporal dynamics: interacting variables: equilibrium, oscillation, chaos

B) drug propagation and control

(adapted from Behrens et. al, 2004)

Figure 8: Drug market might be chaotic

Since a chaotic drug market is interpreted as unpredictable, there is a natural question whether it is possible to control the system to stable cycles or even (low level) equilibrium.
C) economic cycles and chaos

Business Cycles: Kaldor model

Figure 9: Kaldor model: $y$: income, $k$: capital, $I$: investment, $S$: saving
C) economic cycles and chaos

Business Cycles: Kaldor model

Assuming time-independent investment and saving the dynamics is given as:

\[ \dot{y}(t) = a[I(k(t), y(t)) - S(k(t), y(t))] \]  \hspace{1cm} (5a)

\[ \dot{k}(t) = I(k(t), y(t)) - \delta k(t). \]  \hspace{1cm} (5b)
C) economic cycles and chaos

Business Cycles: Kaldor model

Figure 10: Limit cycle behavior of the Kaldor model
Temporal dynamics: interacting variables: equilibrium, oscillation, chaos

D) biological and social epidemics

Figure 11: The standard epidemic model with three sub-populations: $S(t)$ – Susceptible, $I(t)$ – Infectious, $R(t)$ – Recovered
The Actual Data Set

2014 West African Ebola epidemic
Rate of reported cases based on the population of the countries, as of 21 December 2014

The rate is calculated by dividing the number of reported cumulated cases by the population size.
The Actual Data Set

Ebola outbreak

**AFFECTED AREAS**
- Reported cases and deaths
- Area under surveillance

**CUMULATIVE CASES & DEATHS**
Confirmed, probable, and suspected

**CASES & DEATHS BY COUNTRY**
Confirmed, probable, and suspected

Sources: World Health Organization (WHO); UN Office for the Coordination of Humanitarian Affairs (OCHA); Reuters.

Data as of August 6.

Start: 13/08/2014
• How to characterize the changes?

• Difficulties of predictions

• Initial stage of an exponential curve looks linear

• Control strategies
  – to reduce increase to reach inflexion point
  – to reduce increase to get saturation

INFLEXION point: the signature of changing dynamics
The Basic Epidemic Model

• What is happening? "Susceptible" will be converted to "infected"

• What is the velocity of the infection? \( \beta \) determines the "effectivity" of the encounter

• Infection would be ever-increasing: this is not the whole story
The Basic Epidemic Model

- \( S(t) \): represents the number of individuals not yet infected with the disease at time \( t \), or those susceptible to the disease.

- \( I(t) \): the number of individuals who have been infected with the disease and are capable of spreading the disease to those in the susceptible category.

- \( R(t) \): the number of those individuals who have been infected and THEN removed from the disease, due to immunization, quarantine or due to death.

Figure 12: SIR model and its variations
Do you know what Public Health Authorities Do?

They try to influence the $\beta$ and $\gamma$ parameters of the differential equations of epidemic propagation. (stay tuned: two slides later!)

So, please, please, don’t close your eyes!

ASSUMPTION: rate of contact between two groups in a population is proportional to the size of each of the groups concerned $s$

- velocity of change in the number of susceptible ($S$): proportional with the numbers of $S$ and $I$

- velocity of change in the number of infected ($I$): two terms: new infected - removed

- velocity of change in the number of removed: removal rate: proportional with the number of infected.
The Basic Epidemic Model

\[ \frac{dS}{dt} = -\beta SI \]
\[ \frac{dI}{dt} = \beta SI - \gamma I \]
\[ \frac{dR}{dt} = \gamma I \]

- \( \beta \): how effectives are the infected
- \( \gamma \): how effective the removal of the infected (good and bad)
- larger \( \beta \) increases the velocity of the epidemics
- larger \( \gamma \) decreases the velocity of the epidemics
The SIR Model: a typical "solution"

How the quantities of S, I and R are CHANGING in time?
Epidemic is a THRESHOLD PHENOMENON:

There is a threshold quantity which determines whether an epidemic occurs or the disease simply dies out. The threshold is called as the basic reproduction number, often denoted by $R_0$, ($N$ is the total population, initially $S(0)$.

$$R_0 := \frac{\beta N}{\gamma},$$  \hspace{1cm} (6)

The basic reproduction number, $R_0$, is interpreted as the average number of secondary cases caused by a typical infected individual throughout its entire course of infection in a completely susceptible population and in the absence of control intervention. The targets of any control strategies are the $\beta$ and $\gamma$ rate parameters.
A More Advanced Model

"Rivers et al model
(PLOS Currents: Outbreaks:
2014 October/November)

The population is divided into six compartments:

Susceptible (S), Exposed (E), Infectious (I), Hospitalized (F), Funeral (F) - indicating transmission from handling a diseased patient’s body, and Recovered/Removed (R).

Arrows indicate the possible transitions, with the rate parameters.
A More Advanced Model

- Kamal’s simulations
- "boom and bust" dynamics: the **TIME** and **MAGNITUDE** of the maximal value
- I. the temporal course of the epidemics (the paradoxical meaning of "zero infected")
- II. effect of changing the effectiveness of the contact between susceptible and infected
- III. effect of changing the fatal rate
A More Advanced Model
A More Advanced Model

Different trajectories of infected population (I + H + F) with different community contact rate (Bi)

- Bi = 0.0
- Bi = 0.1
- Bi = 0.3
- Bi = 0.6
- Bi = 0.9

Fraction of infected population (I + H + F) vs. Time (days)
A More Advanced Model

Different trajectories of infected population \((I + H + F)\) with different case fatality rate (Delta 1, Delta 2)

- \(\text{Delta 1, Delta 2} = 0.0\)
- \(\text{Delta 1, Delta 2} = 0.3\)
- \(\text{Delta 1, Delta 2} = 0.6\)
- \(\text{Delta 1, Delta 2} = 0.9\)
A More Advanced Model: further possibilities

- More compartments: (say: "good" and "bad" hospital)
- More processes: back to the community from "Temporarily removed"
- Spatial Models
- Interregional Mobility
- Additional 'normal' birth-and-death rate
- Randomness; the world of stochastic models (POSTER!!)
Predictability: Scope and Limits

Is West Africa Approaching a Catastrophic Phase or is the 2014 Ebola Epidemic Slowing Down? Different Models Yield Different Answers for Liberia (Gerardo Chowell, Lone Simonsen, Cécile Viboud, Yang Kuang) November 20, 2014):
Control strategies:
to decrease infection probability
to remove and cure infected (and infectious) subpopulation
BIOLOGICAL and SOCIAL EPIDEMICS: PROPAGATION and CONTROL of INFECTIOUS DISEASES AND IDEAS

**Coupled** Contagion Dynamics of Fear and Disease: Mathematical and Computational Explorations


very important topic: beyond the scope of this presentation
Take home message I

In response to the immediate need for solutions in the field of computational biology against Ebola, The International Society for Computational Biology (ISCB) announces the ISCB Fight Against Ebola Award. ISCB will give out the ISCB Fight Against Ebola Award, along with a prize of $2000, at its July 2016 annual meeting (ISCB ISMB 2016, Orlando, Florida). All computational approaches should include a significant component of Ebola research. In the development of any modern drug, computational biology is positioned to contribute through comparative analysis of the genome sequences of Ebola strains, and 3-D protein modeling. Other computational approaches to Ebola include large-scale docking studies of Ebola proteins with human proteins and with small-molecule libraries, computational modeling of the spread of the virus, computational mining of the Ebola literature, and creation of a curated Ebola database. Taken together, such computational efforts could significantly accelerate traditional scientific approaches.
Take home message II

Mathematical models: A key tool for outbreak response.

• The 2014 outbreak of Ebola in West Africa is unprecedented in its size and geographic range, and demands swift, effective action from the international community
• Understanding the dynamics and spread of Ebola is critical for directing interventions and extinguishing the epidemic
• Mathematical models can clarify how the disease is spreading and provide timely guidance to policymakers
• However, the use of models in public health often meets resistance
  – Public skepticism (and ignorance)
  – Models are often portrayed as arcane and largely inaccessible thought experiments
  – However, the role of models is crucial: they can be used to quantify the effect of mitigation efforts, provide guidance on the scale of interventions required to achieve containment, and identify factors that fundamentally influence the course of an outbreak
Temporal dynamics: interacting variables: equilibrium, oscillation, chaos

A) war dynamics

B) drug propagation and control

C) economic cycles and chaos

D) biological and social epidemics

Figure 13: Summary of some low-dimensional sociodynamic models
Spatiotemporal processes

closed systems: homogenization
open systems: pattern formation

spatial: as urban (racial) segregation (Thomas Schelling: Micromotives and Macrobehavior)

- propagation of matter, energy: Physics, virus: Biology
- propagation of matter: goods, money etc.: Economics
- propagation of drugs: Criminology
- propagation of “information” (knowledge, opinion, gossip, fear, trust) among humans: Behavioral Social sciences

Figure 14: A spatial pattern
”All models are wrong, but some are useful.”

Concept-driven, low-dimensional, single-scale models give insights

Correlation (and brute force) is not enough: we need causal (maybe circular and network causal, but causal) explanations even in the peta-computing age. It is time to combine concept-driven and data-driven models

- to determine governing rules of the dynamics
- to extend to High-dimensional, Multi-Scale, Multi-Level systems
- to obtain realistic parameter values with data mining
- equation-based AND (not XOR) agent-based simulations
Sociodynamics: from concepts to data and back

• The difficulties in social sciences: from Kepler to Newton

• Kepler’s laws: data-driven, inductive, global, integral

• Newton’s laws: data-motivated and concepts driven, deductive, local, differential

• The a priori knowledge of the forcing functions, which basically governs the dynamic evolution of the systems, is difficult: social dynamics is in the post-Keplerian and pre-Newtonian age (and maybe it remains there).
Prediction of emerging technologies based on analysis of the U.S patent citation network

1. General Plan
2. Preliminary Studies
3. Methodology
4. Preliminary Results
5. Conclusions and Plans
General Plan

Conceptual frameworks

• to develop, validate and test a **new technique** about new directions of technological development
• patent citation network
• predictive analytics

Working hypotheses

1. the evolution of the patent citation network reflects (if imperfectly) technological evolution

2. a quantity, the **citation vector**, can be defined appropriately to play the role of a predictor, i.e., to characterize the temporal change of technological fields;

3. clusters of patents, which are the signature of new developmental directions, can be identified based on patterns of similarity in the citations they receive.
General Plan

Technology classification systems

- Very large network (> 4 million nodes between 1975 and 2005)
- Data available electronically (NBER dataset + USPTO)
- USPTO: 450 classes, and over 120,000 patent subclasses
- New classes added; patents can be reclassified
- NBER: 36 sub-categories further lumped into six categories

Computers and Communications, Drugs and Medical, Electrical and Electronics, Chemical, Mechanical and Others
General Plan Evolving clusters

Figure 15: Possible elementary events of cluster evolution. Based on Palla et al 2007
Specific aims

1. to provide a general predictive analytic methodology, which is able to identify structural changes in the patent cluster system and reveal precursors of emerging new technological fields

2. to test and validate the predictive force of the new methodology based on historical examples of new class formation

3. to identify specific mechanisms of the recombination process and formation of new classes

4. to scan the database to identify "hot spots" that may reflect incipient development of new technological clusters
Figure 16: Sections from the in-degree and age based maximum likelihood fitted kernel function for the US patent citation network. Both plots have logarithmic axes. From Gábor Csárdi.
Definition of a predictor for the technological development

Figure 17: Illustration of citation vector calculation in case of four technological categories denoted by the four different colors. The outgoing citations are weighted by the out-degree of their source. The citations originating from the same category (blue in this case) are excluded from the citation vector and the corresponding vector component is set to zero. The received weighted citations are summed and normalized in order to obtain the citation vector.
Methodology

Definition of a predictor for the technological development

1. We select a time point between 1975 and 2007 denoted by $t_1$ and drop all patents that were issued after $t_1$.
2. We compute the citation vector for each patent that remains in the database. We drop patents that have been cited only within their own subcategories.
3. We keep only the patents in some subset of subcategories: $c_1, c_2, \ldots, c_n$ – this step is necessary in order to work with a reasonably sized problem.
4. We compute a similarity matrix based on similarity between patents with respect to their citation distribution as defined above.
5. We apply a clustering algorithm to reveal the functional clusters of patents.
6. We repeat the above steps for several time points $t_1 < t_2 < \cdots < t_n$. 
Methodology

Identification of patent clusters

- to select and test clustering and graph partitioning algorithms to produce sufficiently good results for comparing and validating the clustering results.

- time complexity: an unavoidable trade-off between accuracy and time-consumption.

- the appropriate number of clusters are not known a priori: use hierarchical methods, which do not require that the number of clusters to be specified in advance.

- k-means and the Ward method, which are point clustering algorithms

- graph clustering algorithms: edge-betweenness random walks and the MCL method.
Detection of structural changes in the patent cluster system

- ASSUMPTION: the structure of dendrograms REFLECTS the structural relationships between patent clusters

- In this hierarchy, each branching point is binary and defined only by its height on the dendrogram, corresponding to the distance between the two branches.

- Temporal changes in the cluster structure can be divided into four elementary events: 1) increase or 2) decrease of an existing branching point, 3) insertion of a new and 4) fusion of two existing branching points.

- To find these structural changes, we will identify the corresponding branching points in the dendrograms representing the consecutive time samples of the network and follow their evolution through the time period documented in the database.

- Specifically, potential new classes can be found by comparing the dendrogram structure with the USPTO classification.

- While some of the branching points of the dendrogram are reflected in the current classification structure, there could be significant branches which are not identified by the current system.
Local densities of patents exist in the citation space and can be found with clustering.

Figure 18: Cluster structure of patents in the citation space. Two-dimensional representation of patent similarity structure in the sub-category 11 by using the Fruchterman-Reingold algorithm. Local densities corresponding to technological areas can be recognized by naked eye or identified by clustering methods. The coloring corresponds to the result of hierarchical clustering using Ward method and cut off at the level of seven clusters.
Results

Changes in the structure of clusters reflects technological evolution

Figure 19: Temporal changes in the cluster structure of the patent system. Dendrograms representing the results of the hierarchical Ward clustering of patents in the sub-category 11, based on their citation vector similarity in 1994 (graph A) and 2000 (graph B). The x axis denotes a list of patents in sub-category 11, while the distances between them as defined by the citation vector similarity, are drawn on the y axis. (Patents separated by 0 distance form thin lines on the x axis.) The 7 colors of the dendrogram correspond to the 7 most widely separated clusters. While the overall structure similar in 1994 and 2000, interesting structural changes emerged in this period. The cluster 5 marked with the red color approximately corresponds to the new class 442, which was established in 1997, but was clearly identifiable by our clustering algorithm as early as 1994.
The emergence of new classes: an illustration

Figure 20: An example of the splitting process in the citation space, underlying the formation of a new class. In the 2D projection of the 36 dimensional citation space, position of the circles denote the position of the patents in subcategory 11 in the citation space in three different stages of the separation process (Jan. 1, 1994, Jan. 1, 1997, Dec. 31, 1999). Red circles show those patents which were reclassified into the newly formed class 442, during the year 1997. The rest of the patents which reserved their classification after 1997 are denoted by blue circles. Precursors of the separation appear well before the official establishment of the new class.
Figure 21: Separation of the patents by clustering in the citation space, based on the Jan. 1, 1994 data. **A**: Distribution of the patents issued before 1994 in the subcategory 11, within the 6 official classes in 1997 on the class axis (also marked with different colors) and within the 7 clusters in the citation space. The clustering algorithm collected the majority of those patents which were later reclassified into the newly formed class 442 (red line) into the cluster 5 (both are marked with asterisk). Vice-verse, the cluster 5 contains almost exclusively such patents which were later reclassified. Thus, we were able to identify the precursors of the shaping new class by clustering in the citation space. **B**: The dendrogram belonging to the hierarchical clustering of the patents in the subcategory 11 in year 1994 shows that the branch which belongs to the cluster 5 is the fourth strongest branch of the tree. The coloring here refers to the result of the clustering, thus it is different from the colors in graph A.
Conclusions and Plans

- Patent citation network is a good source of information for making predictions for technological development

- Clustering methods should be tested and validated

- Mechanisms of new class formations will be studied

- Database should be scanned to detect "hot spots" of emerging fields

- (New hopes with a new generation of students)
Towards a Predictive Theory of the US National Budget

An integrated theory of budgetary politics and some empirical tests: The U.S. national budget, 1791-2010

Bryan D. Jones; László Zalányi; Péter Erdi
University of Texas at Austin • Government

We develop a general theory of budgetary politics and examine its implications on a new data set on U.S. government expenditures from 1791 to 2010. We draw on three major approaches to budgeting: decision-making theories, primarily incrementalism and serial processing; policy process models; and path dependency. We show that the incrementalist budget model is recursive and that its solution is exponential growth, and isolate three periods in which it operates in pure form. The equilibrium periods are separated by critical junctures, associated with wars or economic collapse. We assess policy process dynamics by examining the deviations within equilibrium periods. We offer three takeaways: (1) exponential incrementalism is fundamental to a theory of budgeting; (2) disjoint shifts in the level of exponential incrementalism are caused only by critical moments; (3) temporally localized dynamics cause bends in the exponential path, longer returns to the path within budgetary eras, and annual punctuations in budget changes.

©2014, Midwest Political Science Association.

Original language: English (US)
Pages: 561-578
Number of pages: 18
Journal: American Journal of Political Science
Journal publication date: Jan 1 2014
Volume: 58
Issue: 3
DOI: http://dx.doi.org/10.1111/ajps.12088
State: Published
Incrementalist model

- (concept driven) decision making model (bounded rationality)
- players: Requesters and Appropriators
- $R_n$: request in year $n$
- $B_n$: budgetary allocation in year $n$
- external factors are neglected ("closed system")
- tested on data 1947 – 1963 (non-defense agencies)

\begin{align*}
R_n &= \beta B_{n-1} + \xi_n \\
B_n &= \gamma R_n + \zeta_n \\
B_n &= \delta B_{n-1} + \eta_n \\
\delta &= \gamma \beta \\
\eta_n &= \gamma \xi_n + \zeta_n
\end{align*}
Towards a Predictive Theory of the US National Budget

- Exponential Incrementalist model
- self-reinforcing mechanism
- linear positive feedback
- external factors are neglected ("closed system")

\[ \ln B_n = \ln B_0 + \lambda n = A_0 + \lambda n \]  (10)
Towards a Predictive Theory of the US National Budget

- **Disrupted Exponential Incrementalist model**
- trendline: an idea from Ferenc Jánossy
- External factors can disrupt the internally-dominated, closed incremental system
- Critical junctures

Ferenc Jánossy was the most important Hungarian pioneer of surveys on long time series. In the 1960s he devised the famous theory of trendlines, which allowed him to forecast the great world economic
recession of the 1970s a decade in advance.
Towards a Predictive Theory of the US National Budget

- Back to Data!
- Treasury Department, 1791 – 1970
- Office of Management and Budget, 1940 – 2010
- Some inconsistencies: two separate synthetic series, OMB series is more reliable
- The top panel depicts the full historical period, while the bottom panel depicts the post-1950 period. The growth path for US expenditures is exponential, but major deviations occur. Especially noteworthy are the three abrupt ratchets and the distinct curvature after 1980.
- Changes in $A_0$ and $\lambda$
Towards a Predictive Theory of the US National Budget

Statistical Approaches

Method 1
is a smoothing technique applied to the budget series by taking the cumulative sum of the budget values—roughly the numerical integration of these values—allowing us to focus on the main trends in the data. We may think of this as kind of bird’s-eye view of the budget process.

Method 2
is an examination of rates of change instead of budgetary levels, again seeking deviations from the hypothesized exponential path. More specifically, we analyze the logarithm of year-to-year change ratio, $\log(B(t)/B(t-1))$; the derivative of the logarithm of the budget.
- Method 1 for **total budget**

- larger plotted line: the entire historical period,

- inset graph: isolated periods

- four distinct historical periods, separated by major events.

- Exponential "equilibrium" (more precisely constant acceleration) is indicated for three of them, but not for the period between the First and Second World Wars.
Towards a Predictive Theory of the US National Budget

When a department or agency actually spends money: draws money from the Treasury and gives it to someone - that's a budget **outlay**. Federal budget is just a spending plan, outlays represent the real cost.
• Method 1 for **Defense Outlay**

• larger plotted line: the entire historical period,

• inset graph: isolated periods

• five distinct historical periods, separated by major events ...

• President Theodore Roosevelt’s military expansion in around 1900, post-Vietnam withdrawal of military expenditures.
Towards a Predictive Theory of the US National Budget

- **Method 1 for Domestic Outlay**
- Larger plotted line: the entire historical period,
- Inset graph: isolated periods
- There are periods in which exponential growth seems not to be exponentially stable for the domestic budget.
Towards a Predictive Theory of the US National Budget

Fluctuations around the trend

skewness is a measure of the asymmetry of the probability distribution
kurtosis is any measure of the "peakedness" of the probability distribution

Figure 9: Histogram of the Logarithm of Budget Changes for the Full Data Series (1791-2010) for Defense, Domestic, and the Total Budget
- Residuals from Fittings to the Derivative of Log Budget of Stable Periods
- Are critical junctions are part of the general theory?
- If critical junctures are just part of a broader budgetary dynamics that characterizes the whole budgetary series
- then we expect frequency distributions examining only the distributions for the stable periods to resemble these distributions
- Analysis of the residuals in the stable periods.
- residuals are the random adjustments to the general trends
- Compared to the full series analysis presented in Figure 9, the histograms have fewer cases in the tails, and the kurtosis, which assesses punctuations in change data series, is reduced.
- Need for incorporating critical junctures into the theory
- But for some periods the kurtosis remains large, indicating budget punctuations within the stable period
Is the adjustment process stationary? This is true for the middle and late periods for the total budget, but not for the first period. But these deviations do not seem to have effects on the total budget path.
Towards a Predictive Theory of the US National Budget

- basic driver of budget change: self-reinforcing incremental system
- three major periods of budget stability
- critical junctures: changes in the intercept and slope
- after WWII: two stages: from the War to the late 1970s large slope
- followed by deceleration: the last period does not fit well to the general theory
- more general theory is needed to find a social Neptune
Grand conclusions

• Newton’ dynamical model: predictions led to the discovery of Neptune and Pluto

• What to do with this dynamical approach to social sciences?

• Social sciences: the overwhelming majority of data has been evaluated and interpreted in static or equilibrium perspective

• Something should be done with social data deluge

• The a priori knowledge of the rules, which basically govern the dynamic evolution of the systems, is difficult;

• By combining inductive and deductive strategies and using high-performance computing: we hope to predict our social Neptunes and Plutos
Cooperation

Hungarian students:
Kinga Makovi (now: Columbia Univ.)
Peter Volf

Seniors in Budapest:
Zoltán Somogyvári
László Zalányi
Péter Bruck

Seniors in the US:
Katherine Strandburg (NYU)
Jan Tobochnik (Kzoo)
Bryan D Jones (Austin)
Frank Baumgartner (Chapel Hill)

Kalamazoo College students:
Amber C. Salom
Kamalaldin M. Kamalaldin
Hayley Q. Beltz
Raoul R. Wadhwa
Alexander T. Townsend
Takumi Matsuzawa
Louis J. Hochster
Conrad L. Hipkins-Jones
Abhay Goel