Data- & process mining algorithms for the support of early warning systems

Project members: János Abonyi
Krisztina Horák
András Király
Péter Koncz
Tibor Kulcsár
Tamás Ruppert

University of Pannonia
Department of Process Engineering
Veszprém, Hungary
www.abonyilab.com

This research is supported by the European Union and the Hungarian Republic through the GOP-1.1.1-11-2011-0045 project

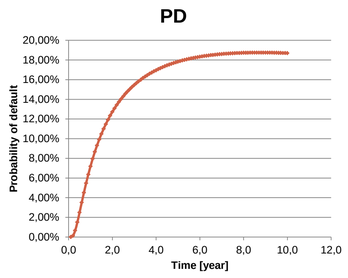
Method

We have developed an early warning framework to predict financial risks of small and medium-sized businesses. As the concept shows below, using our methodology banks will be capable to forecast the probability of default in time, according to firms' transactional data, balance sheets and their position in the business network.

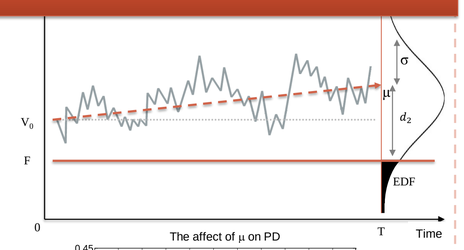
Our concept is the following:

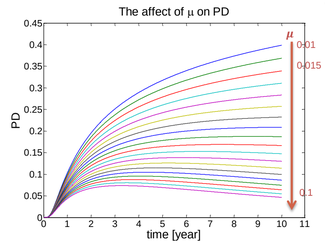
- The Merton model is a structural model that can be used to predict the probability of default (PD) of an individual company. The tool is also used for parameter sensitivity analysis.
- Our stochastic simulator is capable to generate the network of firms and to model the connections and transactions between them. It provides a reliable business network and a transactional log.
- Markov model based credit migration model is identified to give a compact representation of PD.
- A network based epidemic model was developed to evaluate how structural properties of the bussines network infuence the financial risk.
- Using Statistical Process Control (SPC) and GBM methods, the alarm limits of the EWS system can be determined by the combination of Monte-Carlo simulation, Kaplan-Meier estimation and inverz Gauss function.
- Rule-based models for the prediction of defaults can be extracted from logs of events by process mining algorithms.

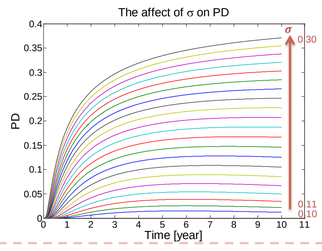
Merton model

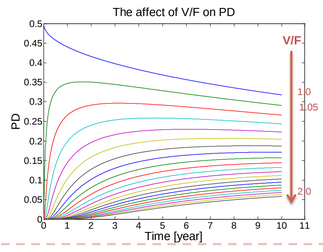


$$P(V_T < F) = P(\ln(V_T) < \ln(F)) = \Phi\left(\frac{\ln\left(\frac{F}{V_0}\right) + \left(\mu - \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right)$$



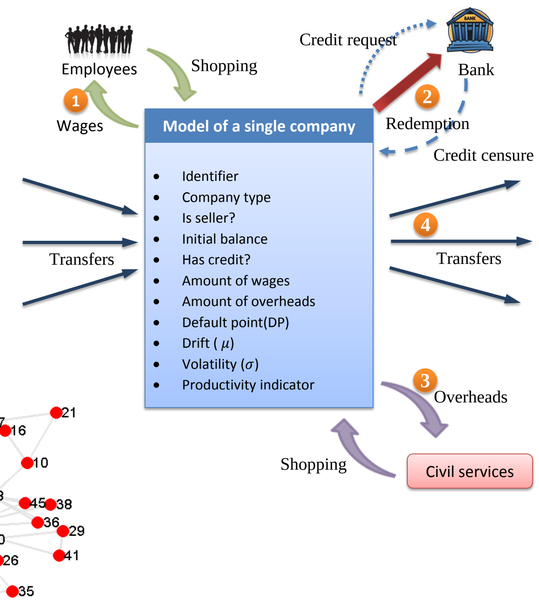


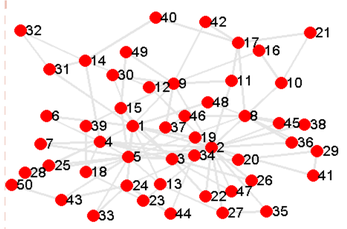




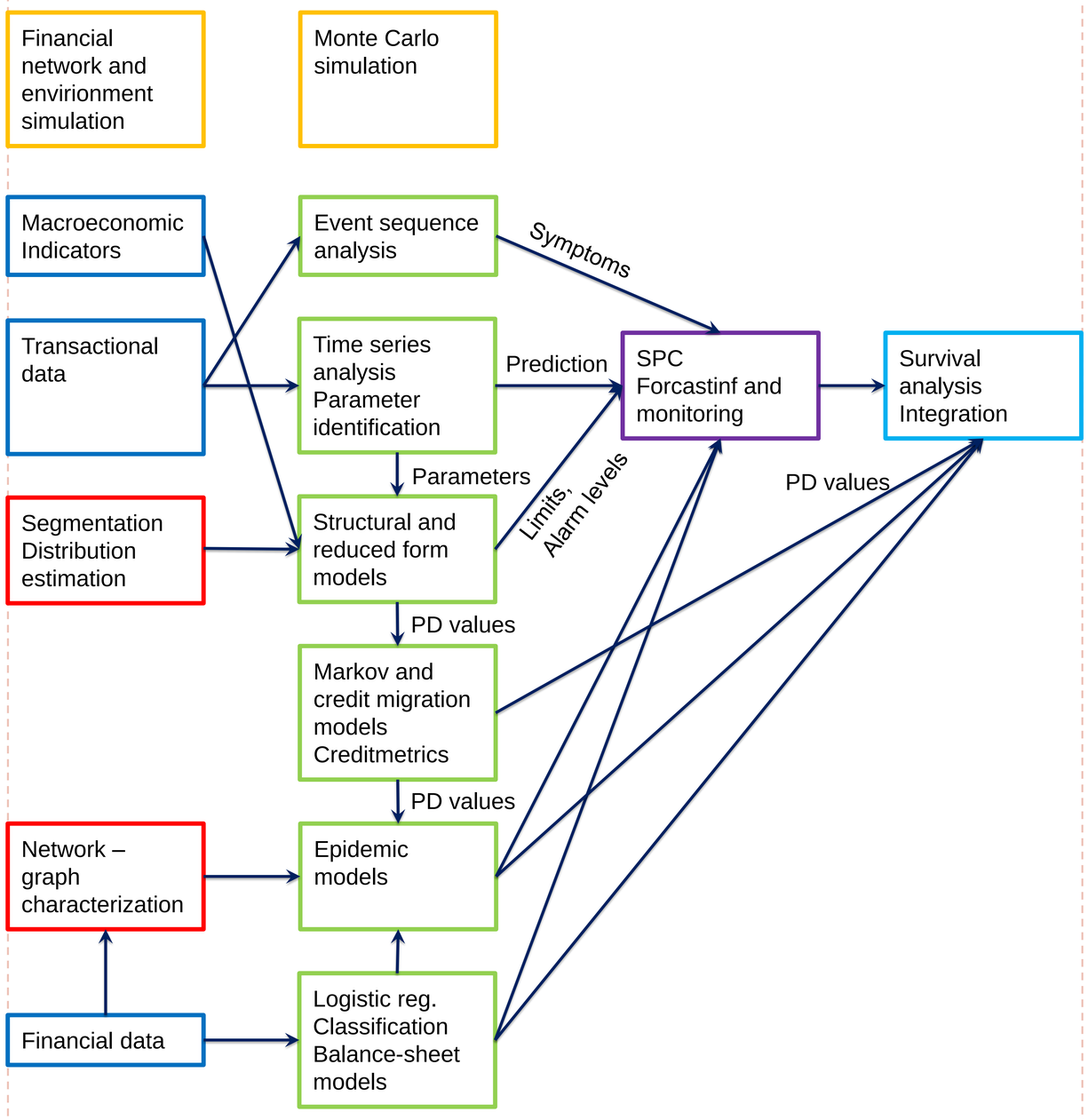
Simulation of a bussiness network

Several graph types
Various debt obligations
Calendar for payments
Tolerance intervals (eg. 90 days for default)
Handy data structure

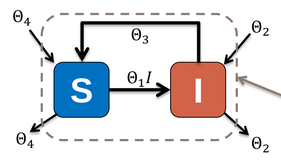




Concept



Epidemic network based model of risk propagation



SIS model for networks

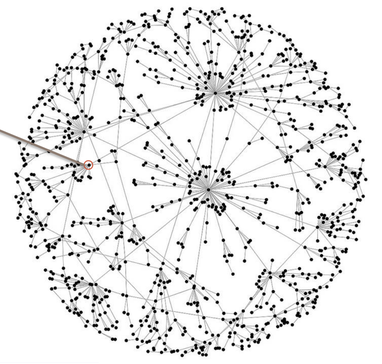
$S_i(t) + I_i(t) = 1$ (Neighborhood)

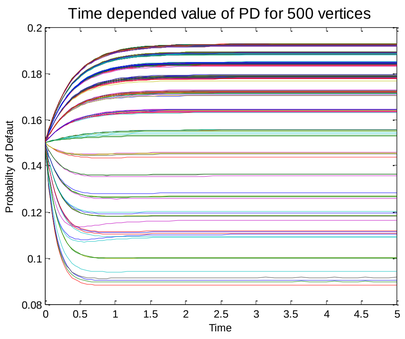
$$\frac{dS_i}{dt} = -\Theta_1 S_i I_i - \Theta_2 S_i \sum_{j=1}^{k_i} I_{ij} + \Theta_3 I_i + \Theta_4 I_i \sum_{j=1}^{k_i} S_{ij}$$

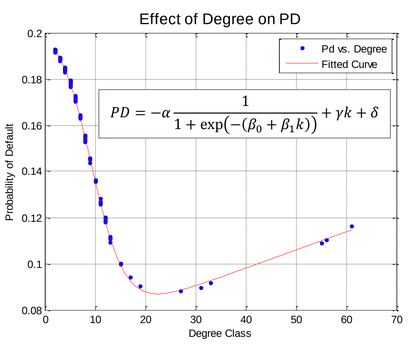
$$\frac{dI_i}{dt} = +\Theta_1 S_i I_i + \Theta_2 S_i \sum_{j=1}^{k_i} I_{ij} - \Theta_3 I_i - \Theta_4 I_i \sum_{j=1}^{k_i} S_{ij}$$

Aggregated form of SIS model for networks

$$\frac{dI_i}{dt} = -\Theta_1 I_i(1 - I_i) - \Theta_2 I_i \sum_{j=1}^{k_i} [G_{i,j}(1 - I_{ij})] + \Theta_3(1 - I_i) + \Theta_4(1 - I_i) \sum_{j=1}^{k_i} G_{i,j} I_{ij}$$

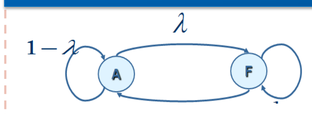




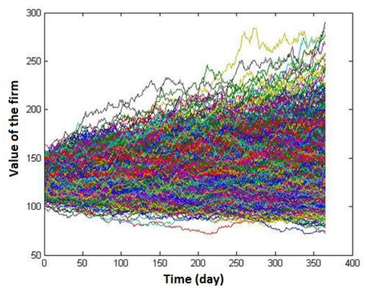


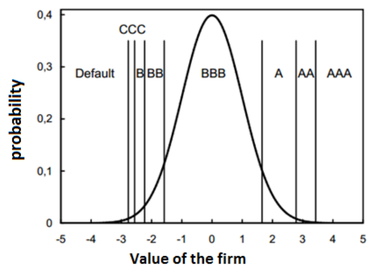
We modified the susceptible-infected-susceptible (SIS) model in scale-free networks.
Each vertex represents an individual of the population and the edges represent the physical interactions among which the infection propagates. We extended the standard SIS structure with neighborhood dependent terms to model the effects of contacts. We solved the coupled differential equation system and analyzed the steady-state solution.
The upper figure demonstrates the solution in the time domain with 0.15 initial condition, while the bottom diagram shows the effect of degree on steady-state PD value.
We studied how degree of the node (number of bissiness partners) influences the PD

Markov & credit migration models

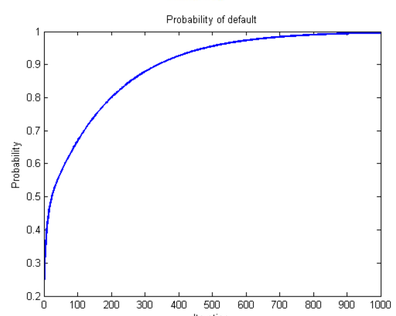


$$P(x_{t+1} = i_{t+1} | X_t = i_t, X_{t-1} = i_{t-1}, \dots) = P(x_{t+1} = i_{t+1} | X_t = i_t)$$

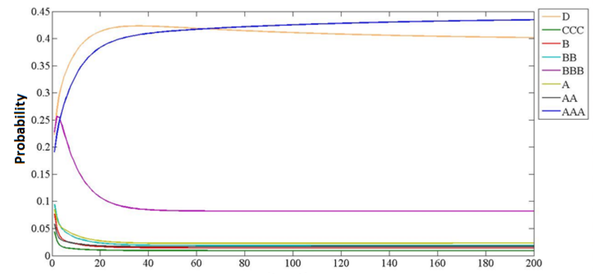




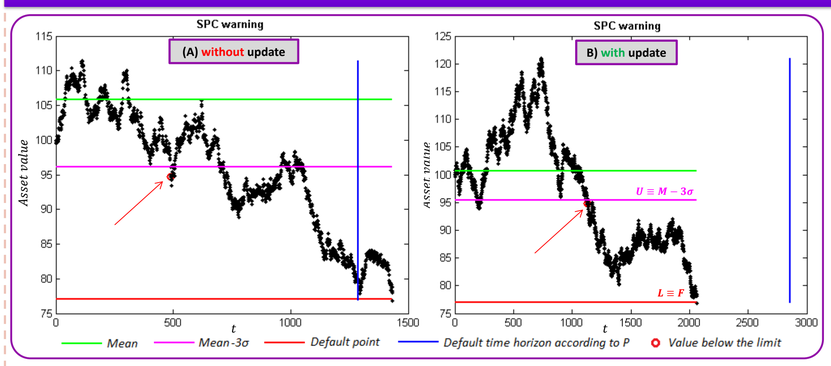
	AAA	AA	A	BBB	BB	B	CCC	Default
AAA	90.81	8.33	0.68	0.06	0.12	0	0	0
AA	0.70	90.65	7.79	0.64	0.06	0.14	0.02	0
A	0.09	2.27	91.05	5.52	0.74	0.26	0.01	0.06
BBB	0.02	0.33	5.95	85.93	5.30	1.17	1.12	0.18
BB	0.03	0.14	0.67	7.73	80.53	8.84	1.00	1.06
B	0.01	0.11	0.24	0.43	6.48	83.46	4.07	5.20
CCC	0.21	0	0.22	1.30	2.38	11.24	64.86	19.79



Extended model

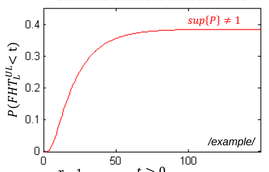


SPC - GBM



(B) with update

GBM between upper limit and lower limit.
GBM hits L before it would hit U.

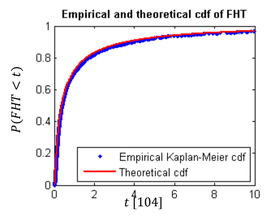


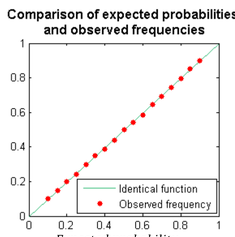
$$g(t) = \left[\frac{L}{V_0}\right]^{\frac{r}{\sigma^2} - \frac{1}{2}} \cdot exp\left\{-\frac{1}{2}\left(\frac{r}{\sigma} - \frac{\sigma}{2}\right)^2 t\right\} \cdot \sum_{n=-\infty}^{\infty} f(t; a_n) \qquad t > 0$$

$$a_n = \frac{1}{\sigma} ln\left(\frac{U^{2n} V_0}{L^{2n+1}}\right) \qquad f(t; a_n) = \frac{a_n}{\sqrt{2\pi t^3}} e^{\frac{-a_n^2}{2t}}$$

A) without update

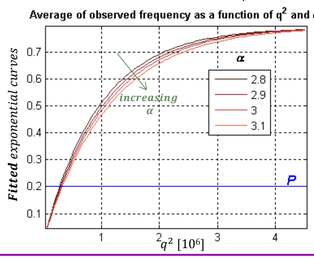
$$P(T_F(V) < t) = \phi\left(\frac{ln\left(\frac{F}{V_0}\right) - \left(r - \frac{\sigma^2}{2}\right)t}{\sigma\sqrt{t}}\right) + exp\left(\frac{2 \cdot ln\left(\frac{F}{V_0}\right) \cdot \left(r - \frac{\sigma^2}{2}\right)}{\sigma^2}\right) \cdot \phi\left(\frac{ln\left(\frac{F}{V_0}\right) + \left(r - \frac{\sigma^2}{2}\right)t}{\sigma\sqrt{t}}\right)$$



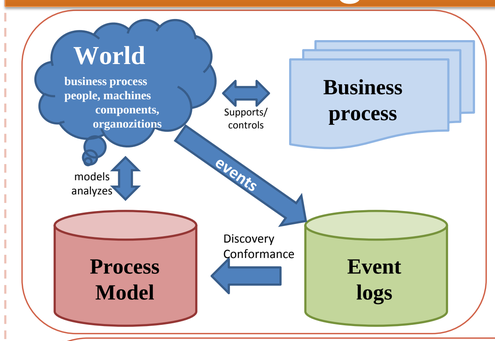


(A) equation for (B) task

$$q := \frac{\sigma}{r} \qquad R(q^2, \alpha, P) = K_{\alpha,P} \cdot \left(1 - exp\left(\frac{-q^2}{\tau_{\alpha,P}}\right)\right)$$



Process Mining



We applied process mining to extract useful information from the logfiles of events.
Evens can transactions, alarms of the SPC system, defaults of neighbours in the business networks, etc.

Bellow the basic concept of process mining is illustrated by the schematic representation of the alpha algorithm.

The algorithm was implemented in MATLAB.

Case ID	Task ID
1	A
2	A
3	A
3	B
1	B
1	C
2	C
4	A
2	B
2	D
5	E
4	C
1	D
3	C
3	D
4	B
5	F
4	D

Alpha algorithm

Direct succession	Causality	Parallel
A > B	A -> B	B || C
A > C	A -> C	C || B
B > C	B -> D	
B > D	E -> F	
C > B	C -> D	
C > D		
E > F		

Adjacencia matrix

	A	B	C	D	E	F
A	0	1	1	0	0	0
B	0	0	1	1	0	0
C	0	1	0	1	0	0
D	0	0	0	0	0	0
E	0	0	0	0	0	1
F	0	0	0	0	0	0

directed graph construction