Supplementary Information: Urban hierarchy and spatial diffusion over the innovation life cycle

Eszter Bokányi[1,2,*], Martin Novák[1,3], Ákos Jakobi[4], Balázs Lengyel[1,2]

[1]ELKH Centre for Economic and Regional Studies; Agglomeration and Social Networks Lendület Research Group, Budapest, H-1097, Hungary

[2]Corvinus University of Budapest; Laboratory for Networks, Technology and Innovation; Budapest H-1093, Hungary

[3]ELKH Wigner Research Centre for Physics; Budapest, H-1121, Hungary

[4]Eötvös Loránd University; Department for Regional Science Budapest, H-1117, Hungary

[*]Corresponding author: bokanyi.eszter@krtk.hu

Supplementary Information 1: The iWiW life cycle

The number of registrations was low in the first few years but iWiW reached a several hundred thousands of people by 2005. The largest Hungarian internet provider company acquired the website in 2006, and the number of registered users grew rapidly from 1.5 to more than 4 millions until December 2008. Moreover, between 2005 and 2010, iWiW was the most frequently visited Hungarian webpage. This is well reflected in Figure 5, that shows the number of registrations over the full life cycle of the website, and confirms that at the peak of iWiW's popularity, the number of registered users jumped to more than 50.000 monthly, and increased further with high variability to a peak about 90.000 invitations per month until the middle of 2007.

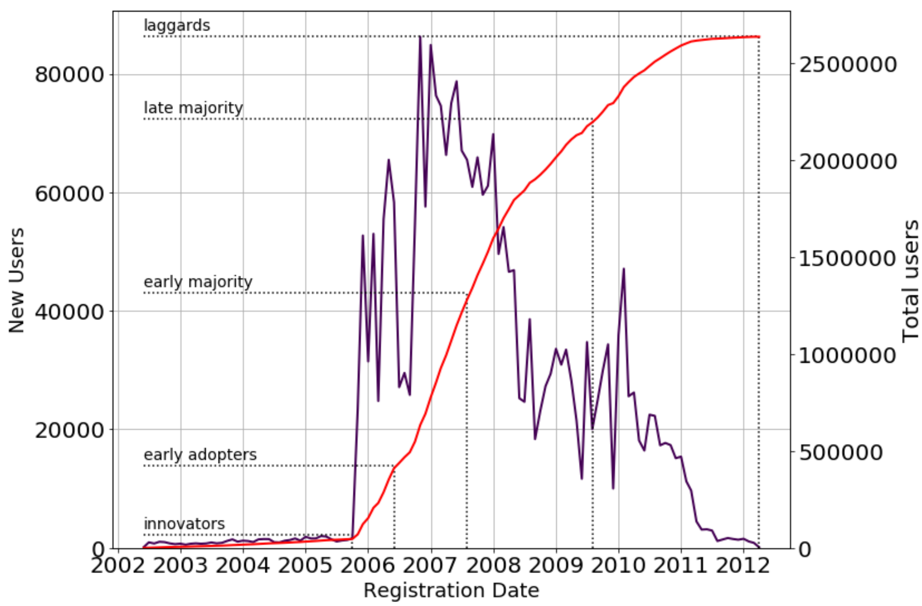

Figure 5: Number of registrations on the iWiW social network by month. Red curve shows the cumulative number of adopters, annotations reflect the adopter categories of Rogers.

The figure shows how early adoption was driven by the age group of 19-29 years, then people of age 8-18 and 30-40 followed. Older age groups between 41-51, 52-62, then 63-73 gradually joined in the adoption process. This means that age structure and age-related social and family contacts of settlements might influence adoption patterns.

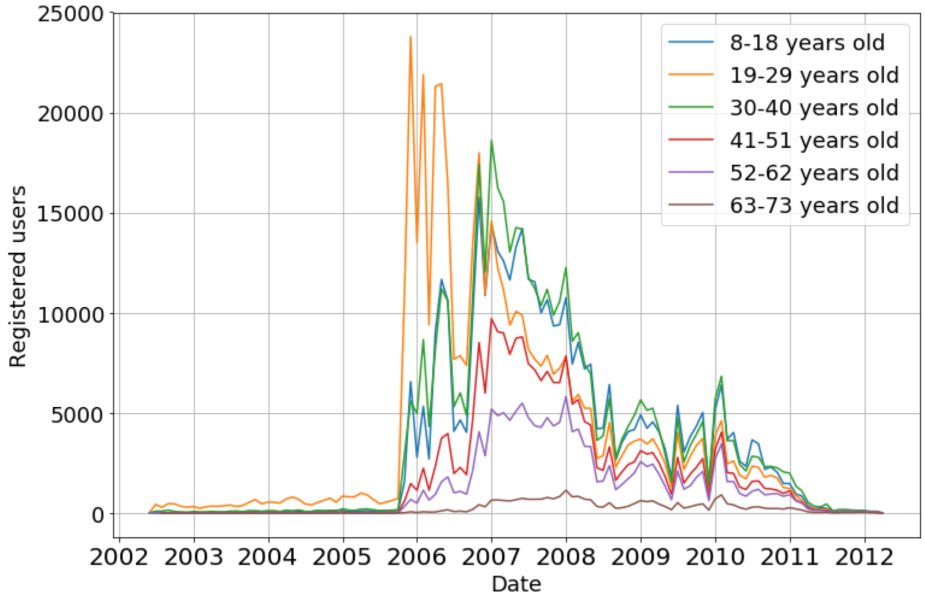

Figure 6: Number of registrations on the iWiW social network by month per age group.

Supplementary Information 2: The share of invitations by diffusion categories

The average distance of invitations declines somewhat in 2006-2007, although the standard deviation of invitation distances marked by the errorbars remains high. However, this distance decrease is not fully in line with the regression coefficient of the $\log d$ term, because here, we do not control for hierarchical patterns.

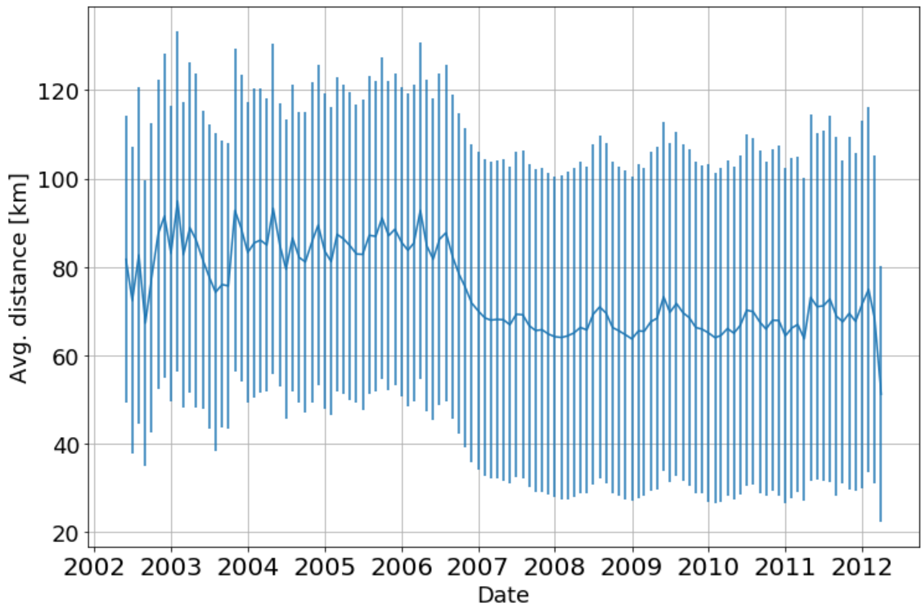

Figure 7: Average distance of invitations in time.

The average population size difference between source and target settlement decreases, that might be connected to the decreasing role of hierarchy in the invitation process.

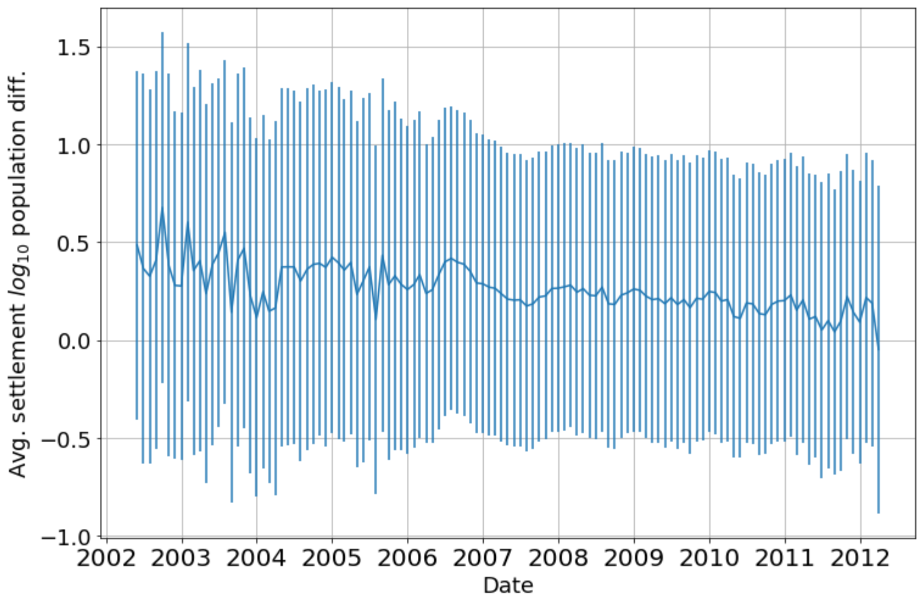

Figure 8: Average population difference of invitations in time.

High share of short-range invitations in the middle of the life cycle is in line with the regression coefficient of the distance term from the negative binomial part of the ZINB regression.

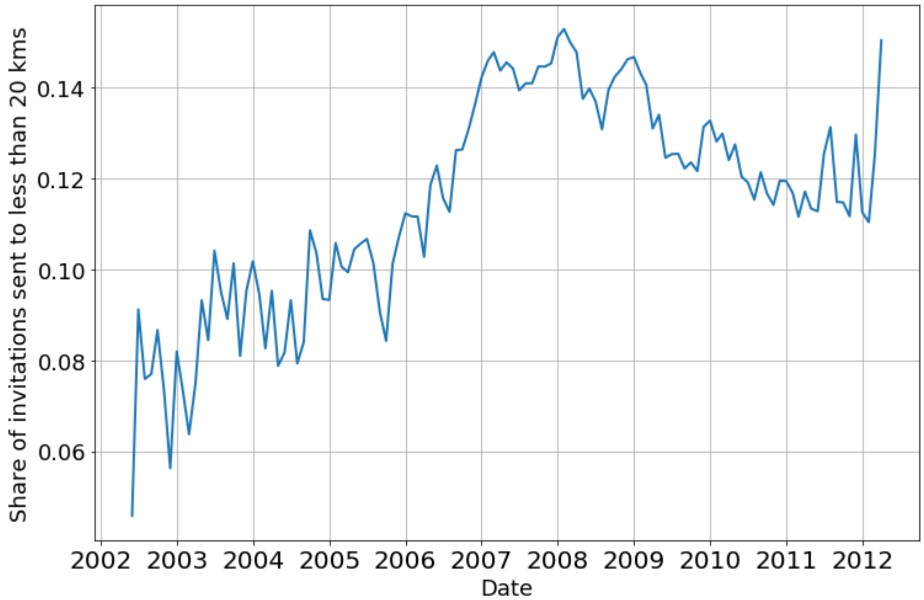

Figure 9: Share of invitations sent to less than 20 km.

The share of invitations sent to a settlement at least 3x smaller than the source settlement is constantly decreasing. In the early and late phases, share of invitations sent into the different direction, to an at least 3x larger settlement than the source settlement is somewhat larger than in the middle of the time range.

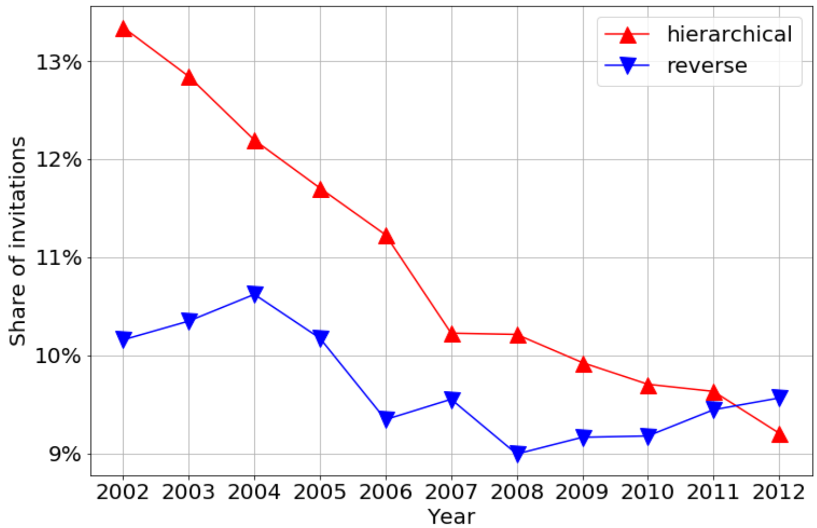

Figure 10: Share of invitations downwards and upwards the hierarchy level (at least 3x population difference in either direction).

The yearly distance distributions show a very pronounced tail for the early years indicating the large weight of long-range invitations in the early stages. High share of short-range invitations indicate the presence of neighborhood diffusion.

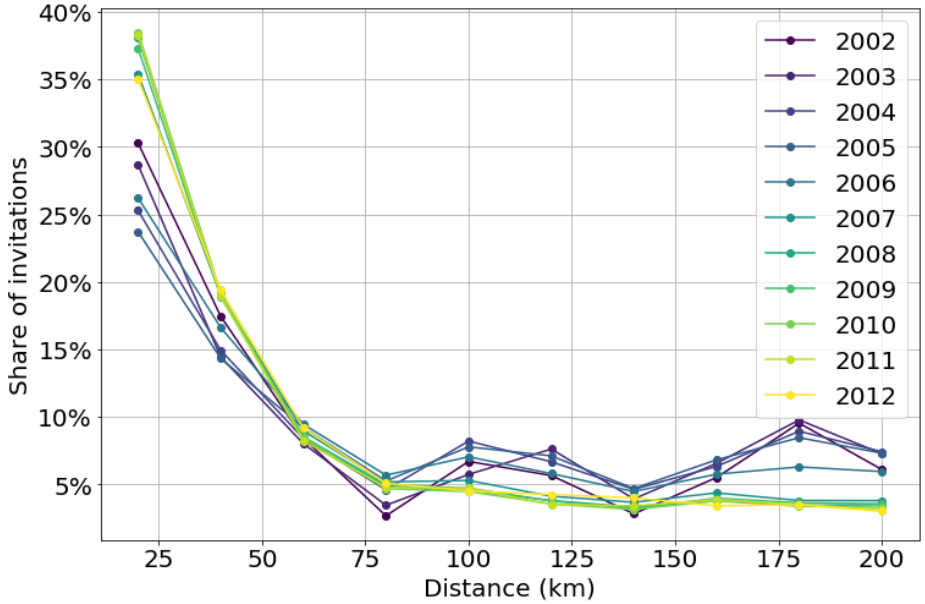

Figure 11: Share of invitations in given distance ranges per year.

The distribution of hierarchy differences in invitations shows a tendency towards hierarchical diffusion, since there are always more invitations sent to smaller settlements (x axis larger than 0), than invitation sent from smaller to larger settlements (x axis smaller than 0). In the early stages, most invitations go between settlements of very different sizes, then similar hierarchy levels tend to dominate the diffusion process.

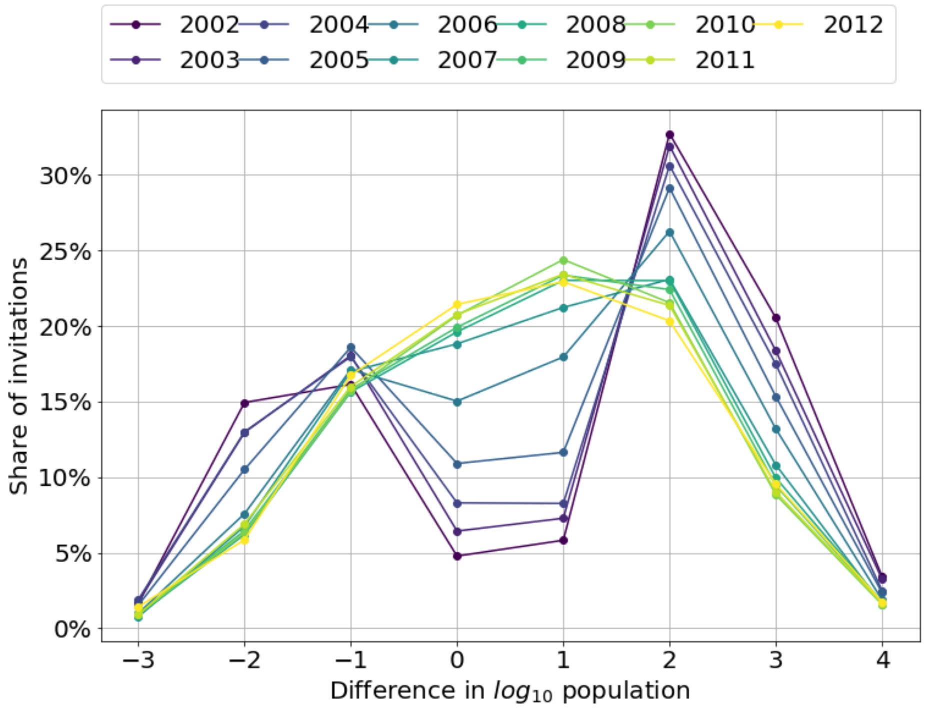

Figure 12: Share of invitations with given log_{10} population difference ranges per year.

Supplementary Information 3: Regression coefficients

In the following two tables, we can find the regression coefficients of the full zero-inflated negative binomial model for all years. Significance levels are marked by stars, standard errors of coefficients are given in brackets. The inflation part corresponds to the estimation of the logit probability term l from equation (2), and the negative binomial part to the estimation of the count process term k from (2). Both terms are estimated by the form

$$\alpha(\log P_S)^2 + \beta \cdot (\log P_S \cdot \log P_T) + \gamma(\log P_T)^2 + \delta \log P_S + \varepsilon \log P_T + \chi \log d_{ST} + C,$$

where α, β, γ, δ, ε, and χ are the coefficients of the variables listed in the regression tables. Note the significance of the second-order terms in the negative binomial part of the model.

	Year	2002		2003		2004		2005		2006		2007	
		Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Model part	Variable												
	Num. obs.	6525466		6525466		6525466		6525466		6525466		6525466	
	$\log \alpha$	-0.568***	(0.176)	-0.832***	(0.147)	-0.895***	(0.126)	-0.792***	(0.0511)	-0.274***	(0.0156)	-0.0661***	(0.0121)
Inflation	Constant	28.52***	(10.84)	32.24***	(5.794)	23.99***	(3.397)	13.05***	(1.433)	-7.983***	(0.78)	-10.57***	(0.665)
	$\log P_S$	-2.452	(1.514)	-2.531***	(0.629)	-2.012***	(0.339)	-0.65***	(0.178)	1.298***	(0.103)	1.635***	(0.0974)
	$\log P_S^2$	0.00491	(0.102)	0.0241	(0.0254)	0.023*	(0.0119)	-0.026***	(0.00704)	-0.111***	(0.00457)	-0.129***	(0.00512)
	$\log P_S \cdot \log P_T$	0.0686	(0.0473)	0.0709***	(0.023)	0.039***	(0.0151)	0.00279	(0.00816)	-0.0171***	(0.00622)	-0.0309***	(0.00636)
	$\log P_T$	-1.777**	(0.845)	-2.201***	(0.55)	-1.215***	(0.369)	-0.69***	(0.157)	1.515***	(0.107)	1.62***	(0.0945)
	$\log P_T^2$	0.0417	(0.0296)	0.0434***	(0.0165)	0.00774	(0.0115)	-0.00393	(0.00505)	-0.119***	(0.00513)	-0.114***	(0.0045)
	$\log d_{ST}$	0.383***	(0.122)	0.487***	(0.071)	0.625***	(0.0511)	0.68***	(0.0236)	0.6***	(0.0124)	0.63***	(0.0105)
Neg. binom.	Constant	15.55	(10.94)	22.22***	(5.611)	17.81***	(3.306)	9.638***	(1.201)	-1.285***	(0.458)	-2.62***	(0.295)
	$\log P_S$	-4.256***	(0.982)	-4.205***	(0.47)	-3.575***	(0.301)	-2.178***	(0.122)	-0.532***	(0.0514)	-0.29***	(0.0355)
	$\log P_S^2$	0.21***	(0.0231)	0.176***	(0.0114)	0.145***	(0.00856)	0.0955***	(0.00358)	0.0506***	(0.0016)	0.0459***	(0.00122)
	$\log P_S \cdot \log P_T$	0.06	(0.0414)	0.115***	(0.0218)	0.113***	(0.0128)	0.104***	(0.00536)	0.0625***	(0.00252)	0.0416***	(0.00186)
	$\log P_T$	-0.778	(0.897)	-1.602***	(0.519)	-1.143***	(0.332)	-0.882***	(0.125)	-0.0645	(0.0511)	0.31***	(0.0354)
	$\log P_T^2$	0.0649***	(0.0187)	0.0661***	(0.0124)	0.0416***	(0.00933)	0.0369***	(0.00358)	0.0231***	(0.00158)	0.0123***	(0.0012)
	$\log d_{ST}$	-1.051***	(0.0526)	-0.882***	(0.039)	-0.787***	(0.0342)	-0.918***	(0.0161)	-1.389***	(0.00789)	-1.605***	(0.00659)

(* = $p < 0.10$, **=$p < 0.05$, ***=$p < 0.01$)

Table 1: Results of the zero-inflated negative binomial regression models for each year (2002-2007).

	Year	2008		2009		2010		2011		2012	
		Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Model part	Variable										
	Num. obs.	6525466		6525466		6525466		6525466		6525466	
	$\log \alpha$	-0.123***	(0.0171)	-0.249***	(0.0218)	-0.315***	(0.0248)	-1.24***	(0.0788)	-2.817*	(1.494)
Inflation	Constant	-7.207***	(0.619)	-6.497***	(0.644)	-6.925***	(0.671)	0.316	(0.952)	13.23***	(3.106)
	$\log P_S$	1.212***	(0.0869)	1.237***	(0.0902)	1.153***	(0.0913)	0.379***	(0.122)	-1.171***	(0.375)
	$\log P_S^2$	-0.094***	(0.00409)	-0.106***	(0.00424)	-0.1***	(0.00419)	-0.0939***	(0.0054)	-0.0482***	(0.0173)
	$\log P_S \cdot \log P_T$	-0.0207***	(0.00577)	0.0077	(0.00578)	0.0124**	(0.00565)	0.0812***	(0.00666)	0.132***	(0.0189)
	$\log P_T$	1.006***	(0.0878)	0.791***	(0.0891)	0.893***	(0.0901)	0.21*	(0.12)	-0.439	(0.388)
	$\log P_T^2$	-0.0809***	(0.00413)	-0.0821***	(0.00418)	-0.0864***	(0.00413)	-0.0858***	(0.00531)	-0.0886***	(0.0174)
	$\log d_{ST}$	0.657***	(0.0123)	0.723***	(0.0139)	0.739***	(0.0155)	1.087***	(0.0249)	1.208***	(0.0791)
Neg. binom.	Constant	-1.812***	(0.31)	-1.763***	(0.349)	-2.677***	(0.369)	-1.183	(0.725)	9.663***	(3.017)
	$\log P_S$	0.00134	(0.0374)	0.0804*	(0.0416)	0.213***	(0.0438)	0.0565	(0.0809)	-1.367***	(0.318)
	$\log P_S^2$	0.031***	(0.00134)	0.0222***	(0.0015)	0.0145***	(0.00158)	-0.00572**	(0.00285)	0.0211*	(0.0116)
	$\log P_S \cdot \log P_T$	0.034***	(0.00194)	0.0386***	(0.0021)	0.0362***	(0.00219)	0.0748***	(0.00358)	0.141***	(0.0114)
	$\log P_T$	-0.096**	(0.0387)	-0.223***	(0.0428)	-0.144***	(0.0445)	-0.442***	(0.0805)	-1.447***	(0.322)
	$\log P_T^2$	0.0301***	(0.00139)	0.0318***	(0.00154)	0.0275***	(0.0016)	0.0143***	(0.00279)	0.0218*	(0.0117)
	$\log d_{ST}$	-1.566***	(0.00909)	-1.476***	(0.0111)	-1.442***	(0.0131)	-1.017***	(0.0212)	-0.709***	(0.066)

(* = $p < 0.10$, **=$p < 0.05$, ***=$p < 0.01$)

Table 2: Results of the zero-inflated negative binomial regression models for each year (2007-2012).

Supplementary Information 4: Goodness-of-fit assessment

Year	TP	FP	FN	TN	Specificity	Sensitivity	RMSE	m	m model	f_s	f_s model
2002	0.000029	0.000015	0.000049	0.999907	0.999985	0.367906	0.0283	2.622309	3.207746	0.000009	0.000006
2003	0.000051	0.000033	0.000113	0.999803	0.999967	0.312150	0.0428	2.987850	4.060219	0.000015	0.000013
2004	0.000082	0.000060	0.000192	0.999666	0.999940	0.298883	0.0610	3.064246	3.891046	0.000024	0.000021
2005	0.000529	0.000373	0.001048	0.998050	0.999626	0.335374	0.3761	3.920505	5.676975	0.000120	0.000096
2006	0.003560	0.002296	0.004804	0.989340	0.997684	0.425677	11.4067	4.791220	10.930057	0.000856	0.000732
2007	0.005420	0.003823	0.007167	0.983590	0.996129	0.430592	14.5399	4.135759	8.752259	0.001310	0.001074
2008	0.003405	0.002515	0.005715	0.988365	0.997461	0.373404	3.7908	2.894744	5.139662	0.000761	0.000568
2009	0.002358	0.001772	0.004508	0.991362	0.998216	0.343451	1.6656	2.463008	3.977552	0.000478	0.000341
2010	0.001965	0.001487	0.004114	0.992434	0.998504	0.323284	1.0236	2.283835	3.424912	0.000388	0.000263
2011	0.000346	0.000287	0.001362	0.998006	0.999712	0.202567	0.0850	1.568390	1.977966	0.000045	0.000030
2012	0.000012	0.000006	0.000154	0.999828	0.999994	0.070175	0.0139	1.103416	1.310345	0.000000	0.000000

Table 3: Goodness-of-fit measures of the ZINB models for the different years. TP=True Positive, FP=False Positive, FN=False Negative, TN=True Negative rates of predicting whether invitation exists between a settlement pair. Specificity = TN / (TN + FP), Sensitivity = TP / (TP + FN), RMSE is the root mean squared error of invitation count predictions, m is observed mean invitation count, m model is predicted mean invitation count, f_s is observed fraction of invitations out of total settlement pairs with at least 5 invitations, f_s model is predicted fraction of invitations out of total settlement pairs with at least 5 invitations.

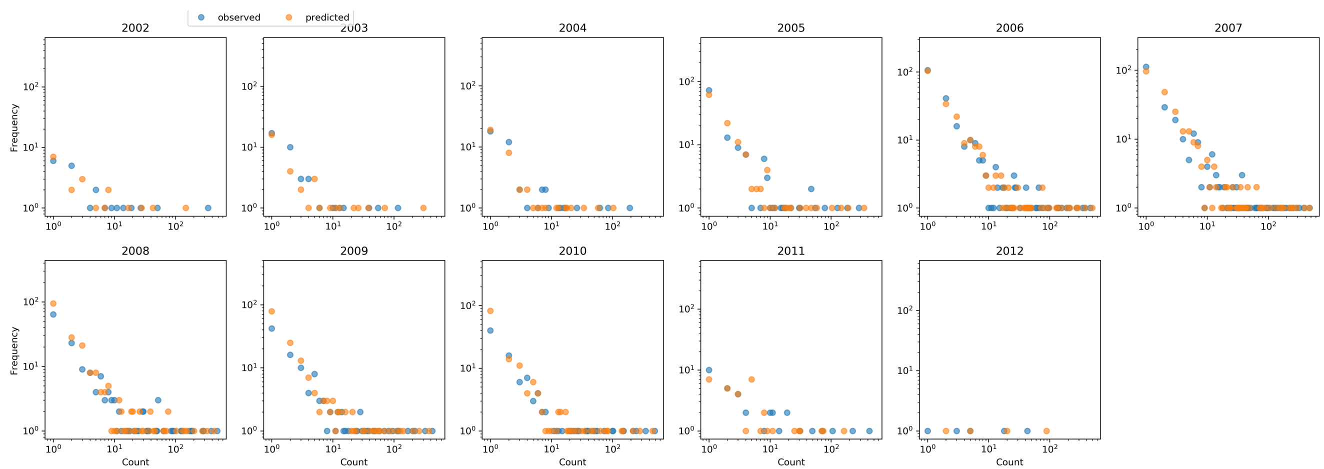

Figure 13: Distribution of count values in the data and ZINB predicted values in each year.

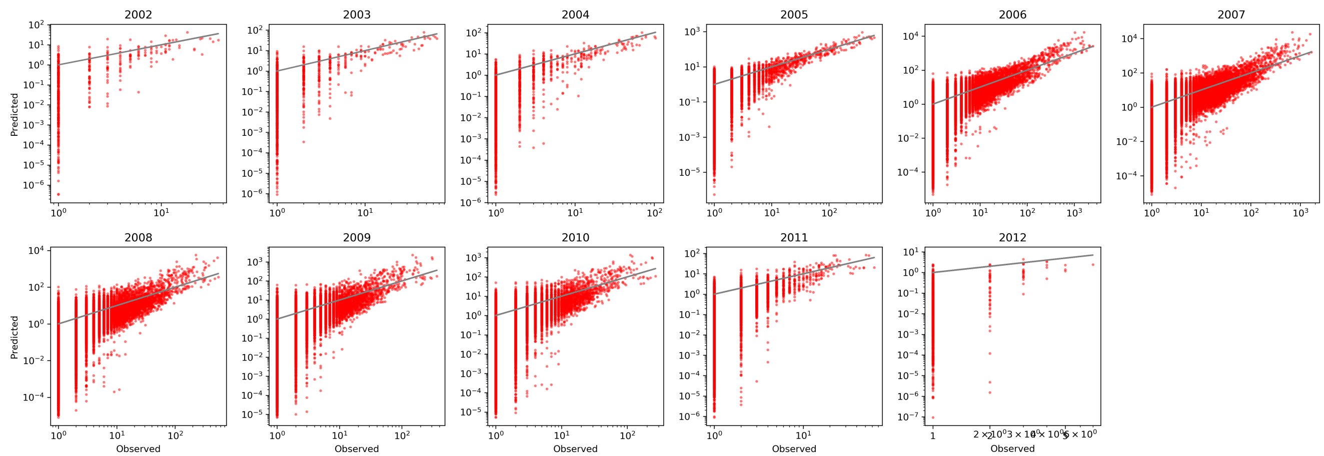

Figure 14: Observed vs. predicted count values in each year, 0 observed counts omitted for better readability.

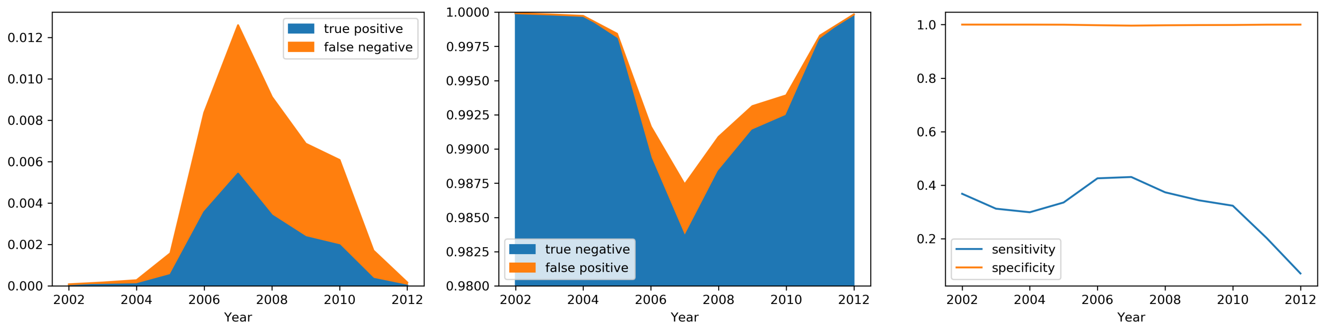

Figure 15: Investigating model performance with respect to predicting settlement pairs between with there are some / there are no invitations.