STUDY 1: ALMOST WIKIPEDIA

Project List

Project	Total Participants	Total Articles
Interpedia	400	<50 (?)
TDEP	1 (?)	5
Everything2	50,000+ (?)	500,000+
h2g2	5,000+	13,000+
TheInfo	20 (?)	50 (?)
Nupedia	2,000+ (?)	24
GNE	300+	3-4 "test" articles
Wikipedia	500,000+	2,000,000+

Table 1: List of OCEPs started in and before January 2001. Details of the size of the projects in total contributors and total articles are shown. These include either the total size over the life of the project or, for the projects that continue today, the total number in December 2010.

Results Summary

Project	P1: Familiar Goal	P2: Low Barriers	P3: Low Social Ownership
Interpedia	No	No	No
TDEP	Yes	No	No
GNE	Yes	No	No
Everything2	No	Yes	No
h2g2	No	Yes	No
Nupedia	Yes	No	No
TheInfo	No	Yes	Yes
Wikipedia	Yes	Yes	Yes

Table 2: Dichotomous codes for each encyclopedia project for each of the three propositions described in the result. A code of "Yes" suggests that there was strong support for that theme in the data associated with the project while "No" suggests there was not strong support.

Results Synthesis

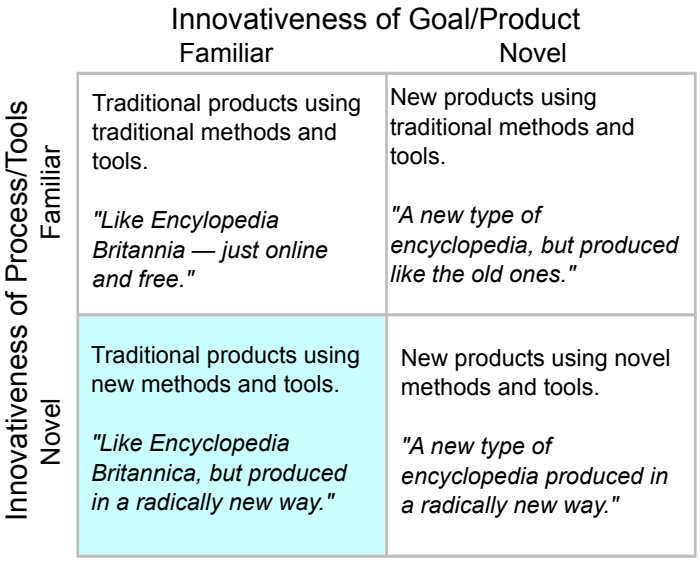

Figure 1: Representation of a theoretical design space in which peer production projects vary between high and low levels of innovation in their goals and products (columns) and processes and tools (rows). I propose that projects in the bottom-left shaded quadrant will be most effective.

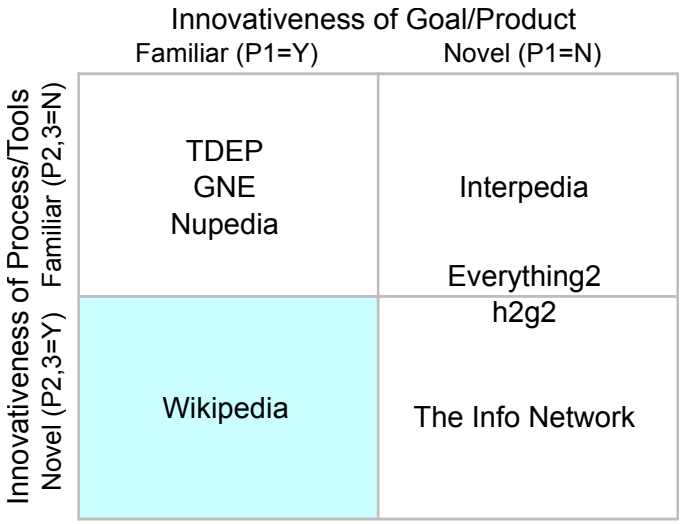

Figure 2: Two-by-two table adapted from Figure 1. Propositions are added to the labels on the axes and the names of OCEPs are placed onto the grid based on their coding as described in Table 2.

STUDY 2: THE REMIXING DILEMMA

Measures and Descriptive Statistics

Variable	N	Mean	SD	Min	Max
Dependent Variables					
Remixes > 0 times in 1 yr. ($remixed_p$)	536245	0.07	0.26	0	1
Remixes within 1 yr. ($remixes_p$)	536245	0.15	1.78	0	658
Edit Distance (Mean) ($distance_p$)	37512	85.57	397.66	0	21970
Question Predictors					
Number of blocks ($blocks_p$)	536245	99.60	476.19	0	196509
User's cumulative views ($userviews_{up}$)	536245	1563.59	5546.90	0	197844
Remix status ($isremix_p$)	536245	0.18	0.38	0	1
Controls					
User age in years (age_{up})	523092	17.57	11.62	4	74.75
Account age in months ($joined_{up}$)	536245	4.79	7.18	0	45.43
User is Female ($female_u$)	536222	0.37	0.48	0	1
Blocks per sprite ($blocks/sprites_p$)	536245	11.82	22.75	0	3111.50
Views within 1 yr. ($views_p$)	536245	13.57	69.90	0	4977

Table 3: Summary statistics for variables used in our analysis. Measures with the subscript p are measured at the level of the project while measures with the subscript u are measured at the level of the user.

Model

Providing tests of Hypotheses 1A-C about generativity, our first two models consider generativity in the full dataset of 523,069 projects shared in our window of data collection for which we have complete information.[1] In our first and more conservative test, Model 1, we use logistic regression to model the likelihood of a project being remixed at least once on our sets of predictors and controls:

$$\text{logit}[\text{P}[remixed_p]] = \beta + \beta \log blocks_p + \beta \log blocks_p^2 + \beta \log userviews_{up} + \beta isremix_p + \beta age_u + \beta joined_{up} + \beta female_u + \beta \log blocks/sprites_p + \beta \log views_p + \beta(\log blocks_p \times isremix_p) + \beta(\log blocks_p^2 \times isremix_p)$$

Model 2 also tests Hypotheses 1A-C using our second measure of generativity: the count of remixes of each project in the first year. It is otherwise identical to Model 1. Poisson regression is frequently used for count dependent variables but, as is common with counts, there is an over-dispersion of zeros in the number of times a project has been remixed. To address this overdispersion, we use a negative binomial regression strategy that estimates the right side of the equation in the model above on the count of *remixes*.

To test Hypotheses 2A-C about originality, we begin with a reduced dataset that consists of the subset of 36,722 projects which were remixed at least once after being shared, and for which we have the creator's age and gender data. The right side of Model 3 is, once again, identical to that of Model 1 shown above. The left side corresponds to the mean Levenshtein distance of every remix of the antecedent project. Because *distance* is a count and, like *remixes*, is overdispersed, we once again forgo Poisson regression in favor of a negative binomial count model.

[1] We omit 13,176 projects for which we are missing age or gender data.

Results

	Generativity		Originality
	Model 1 (P[*remixed*])	Model 2 (*remixes*)	Model 3 (*distance*)
(Intercept)	−5.070***	−5.045***	2.437***
	(0.031)	(0.029)	(0.053)
log *blocks*	0.525***	0.374***	−0.028
	(0.016)	(0.016)	(0.027)
log $blocks^2$	−0.037***	−0.035***	0.052***
	(0.002)	(0.002)	(0.003)
log $userviews_{up}$	0.023***	0.002	−0.041***
	(0.003)	(0.003)	(0.005)
is.remix	0.786***	0.426***	−1.035***
	(0.045)	(0.045)	(0.071)
age	0.000	0.007***	0.006***
	(0.001)	(0.001)	(0.001)
joined	−0.006***	−0.006***	0.005***
	(0.001)	(0.001)	(0.001)
female	−0.003	0.106***	−0.348***
	(0.012)	(0.012)	(0.021)
log *blocks*/*sprites*	−0.517***	−0.375***	0.289***
	(0.012)	(0.012)	(0.018)
log $views_p$	0.840***	1.028***	0.153***
	(0.006)	(0.006)	(0.009)
log *blocks* × *isremix*	0.318***	0.303***	0.160***
	(0.025)	(0.026)	(0.040)
log $blocks^2$ × *isremix*	−0.045***	−0.032***	−0.002
	(0.003)	(0.003)	(0.005)
θ		0.265***	0.301***
		(0.003)	(0.002)
N	523069	523069	36722
AIC	219860.275	307810.178	313334.551
BIC	220396.313	308390.886	313777.130
log L	−109882.137	−153853.089	−156615.276

Standard errors in parentheses; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 4: Model 1 is a logistic regression model of the likelihood of a project being remixed within one year. Model 2 is a negative binomial regression model of a count of the times a project will be remixed within a year. Both use the full dataset of projects ($N = 523,069$). Model 3 is a negative binomial regression model of a count of the mean edit distance for all projects remixed within a year of being shared ($N = 36,722$).

Prototypical Plots

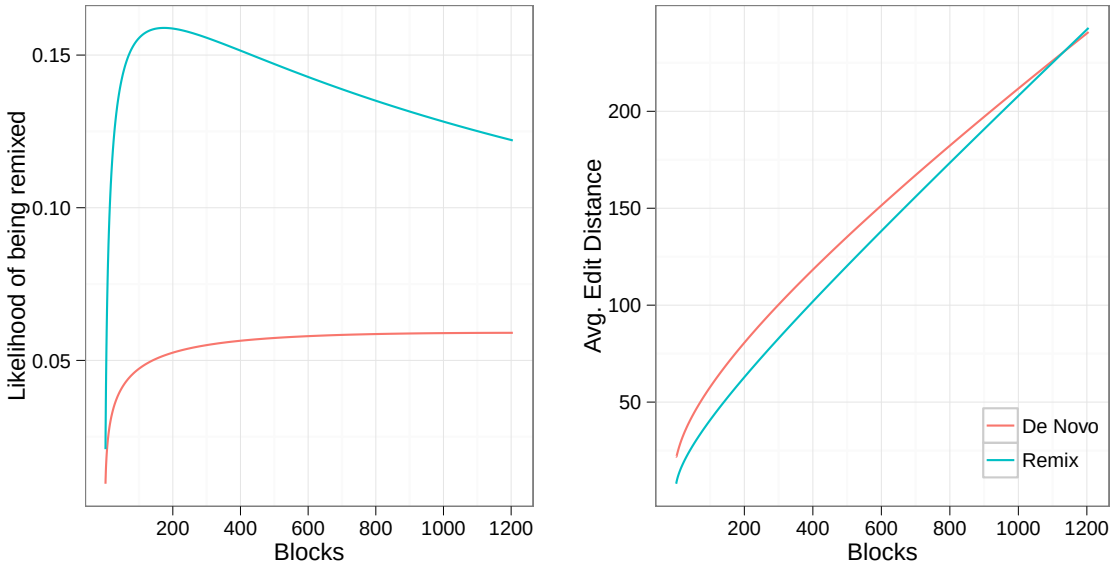

Figure 3: Two plots of estimated values for prototypical projects. Panel 1 (left) display predicted probabilities of being remixed as estimated by Model 1. Panel 2 (right) display predicted edit distances as estimated by from Model 3. Both models show predicted values for both remixes and *de novo* projects from 0 to 1,204 blocks (99^{th} percentile).

STUDY 3: LABORATORIES OF OLIGARCHY?

Descriptive Statistics

Variable	Minimum	Median	Mean	Maximum	SD
Edits	644	13438	53306	2303248	161652
Pages	183	3167	11152	1270640	53025
Editors	69	218	787	68222	3457
Reverted Edits	0	285	1441	122950	5886
Administrators	0	7	11	247	18
Age (Months)	6	46	50	74	11
Project Edits	0	55	622	59726	3224
Experienced User Edits	418	12270	49020	2020925	149606

Table 5: Summary statistics for all of the wikis included in our analysis. ($n = 683$)

Model and Analytic Strategy

Following Singer and Willett (2003), we use hierarchical linear models as a multilevel model for change and fit random intercepts for each wiki to cluster within-wiki variance in a compound error covariance structure.

Each of our models is fit with a measure of oligarchy as its dependent variable and each model corresponds to one of our hypotheses. In Model 1 (M1), we use a multilevel logistic regression to estimate the probability of a new administrator being added. Models 2 and 3 are hierarchical linear models on different dependent variables: (M2) is the log-transformed number of edits to administrative pages by administrators, and (M3) is the log-transformed number of reverts of experienced users by administrators. We use a base model in which every variable is measured at the level of the wiki week and which includes a set of controls as well as our compound error term:

$$Y = \beta accounts + \beta \ln week + \beta \ln week^2 +$$
$$\beta \ln pages + \beta \ln admins + \beta \ln edits + [u + \epsilon]$$

M2 estimates log edits by administrators on administrative pages and adds a control for the total amount of such activity (*ln proj-edits*). M3 estimates the log number of administrator reverts of experienced contributors' edits and includes a control for the total number of these edits (*ln expr-edits*).

Plots from Example Wiki

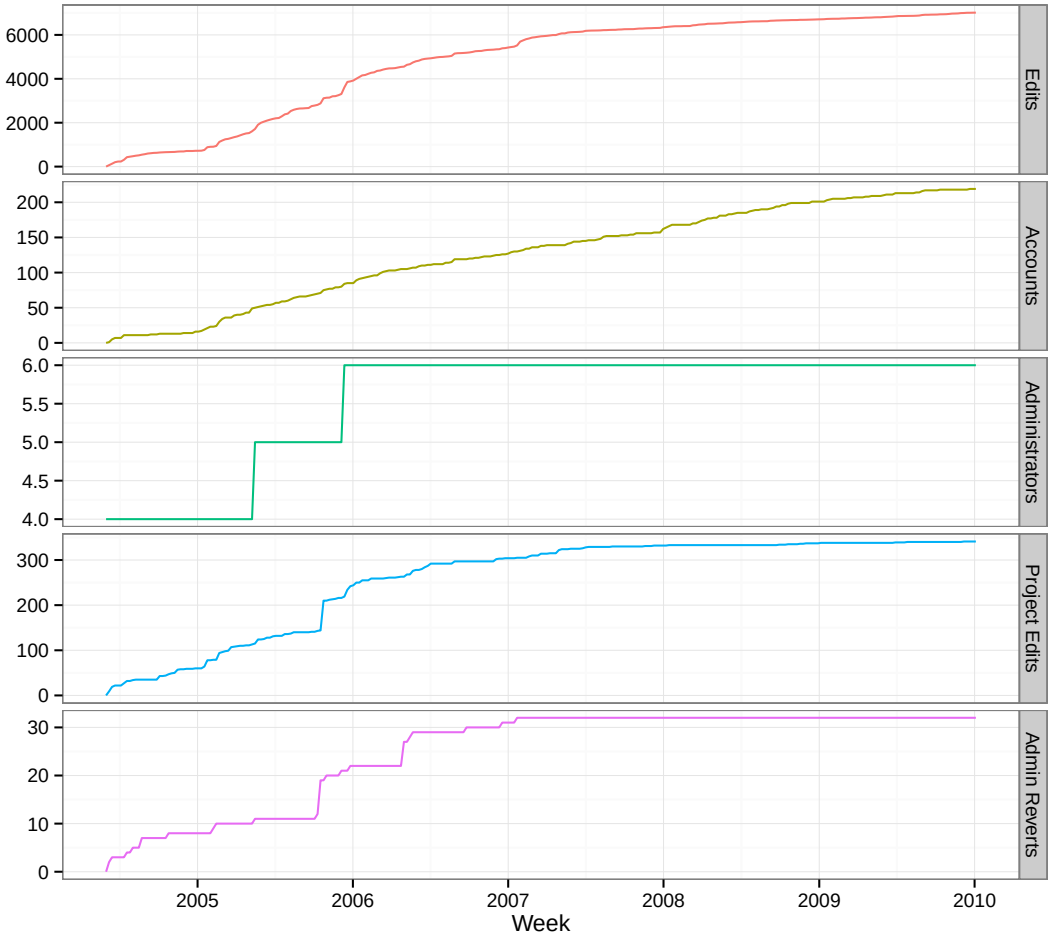

Figure 4: Cumulative plots of covariates for *Seattle Wiki*, a collaborative website for information about Seattle and one of the online communities in our dataset.

Fitted Regression Models

	M1	M2	M3
(Intercept)	—4.108***	—0.138***	—0.029*
	(0.081)	(0.011)	(0.015)
week	—0.006***	0.000***	—0.002***
	(0.001)	(0.000)	(0.000)
$week^2$	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)
ln $accounts_{total}$	—0.210***	0.025***	0.045***
	(0.031)	(0.002)	(0.002)
ln $editors_{week}$	0.116**	—0.036***	0.125***
	(0.038)	(0.002)	(0.003)
ln $pages_{total}$	—0.763***	—0.009***	—0.021***
	(0.023)	(0.001)	(0.002)
ln $admins_{total}$	0.666***	0.070***	—0.010
	(0.035)	(0.005)	(0.006)
ln $edits_{week}$	0.996***	0.010***	—0.111***
	(0.022)	(0.001)	(0.005)
ln $proj\text{-}edits_{week}$		0.608***	
		(0.002)	
ln $expr\text{-}edits_{week}$			0.178***
			(0.004)
Log Likelihood	-11034.750	-38022.642	-103351.946
Num. obs.	146858	118994	146858
Num. groups: wiki	683	554	683
Variance: 1|wiki	0.325	0.034	0.080
Variance: Residual	1.000	0.109	0.234

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6: Table of fitted multilevel regression models. The unit of analysis in each case is the wiki week. *M1* is a logistic regression regression model of the probability that a wiki will add a new administrator during a week. *M2* is a linear model predicting the logged number of edits made by administrators on administrative "project" pages controlling for total edits to these pages. *M3* is a linear model predicting the logged number of reverts of edits by experienced editors by administrators controlling for the number of edits by experienced editors.

Plots of Estimated Values for Prototypical Wikis

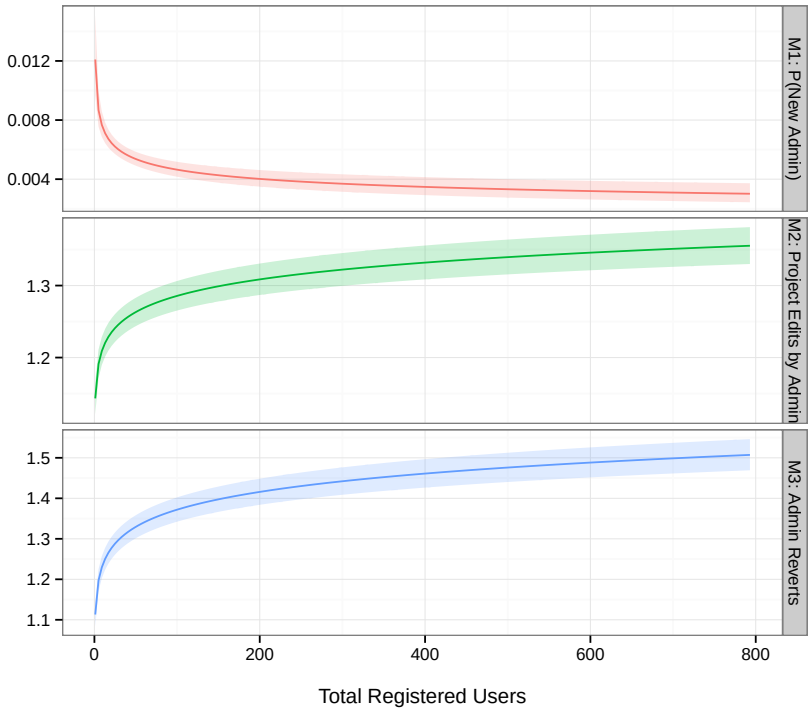

Figure 5: Plots showing predicted values from our models for wikis with varying number of accounts holding all other variables at sample medians. The graph also includes 95% confidence intervals for the marginal effects using the methods and tools described in Fox (2003). All outcome variables are measured in "per week" units.