Website Changes and User Behavior

Using Panjiva Data to Examine Code Changes

John J. Wang

14.27 Final Presentation

December 1, 2012

The Silicon Valley Mindset

- "Move fast and break things." – Facebook
- "The only constant is change itself." – Heraclitus
- "Pick a movement, pick a revolution, and join it." – Jack Dorsey

Background and Previous Research

- Academia has little to say on code changes and user behavior
- Most of the data is hidden away in large tech companies
- Although these companies probably run experiments, results aren't necessarily made public

Lingering Questions

- Do users tend to respond favorably to website changes?
- How do users react to different types of change?
- How do different characteristics of users affect their reactions to change?

Panjiva, Inc.

- `http://www.panjiva.com`
- Acts as a medium for buyers and suppliers of manufactured goods
- Parses government import and export data for unbiased shipment information on suppliers
- Example: Home Depot finding a wrench factory

Summary Statistics

Table: Panjiva Overview

Total Users	121,653
Subscribing Users	2,985
Monthly Site Visits	903,426
Monthly Unique Visitors	762,723
Average Pages per Visit	1.99
Average Visit Duration	1 min 18 sec

Event and Activity Logs

- Event Logs contain records of all user actions 124 million records
- Activity Logs contain records of actions of users who are registered with Panjiva 13 million records

Commit Statistics

Table: Overall Commit Statistics - 11/25/2012

Active Days (at least 1 commit)	1,983
Total Current Files	20,901
Total Lines of Code	1,313,235
Total Lines of Code Added	3,989,295
Total Lines of Code Removed	2,676,060
Total Commits	29,924
Total Authors/Developers	33

Commit Frequency by Time of Day

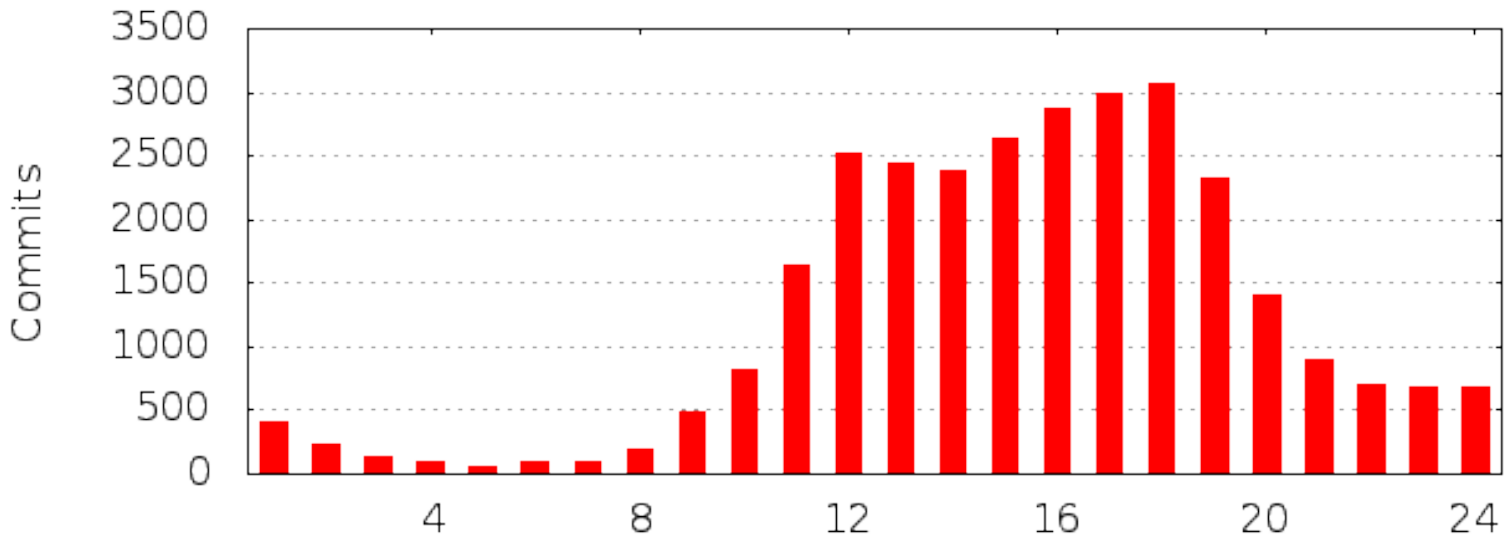


Figure: Number of Commits by Hour of the Day

Specification

$$y_t^i = c_0 + \vec{\gamma}^T \vec{M}_t + \vec{\beta}^T \vec{\chi}_t + \epsilon_t \quad (1)$$

- t indexes day
- y_t^i corresponds to ith metric of user activity on day t
- $\vec{M}_t$ corresponds to a vector of covariates that represent changes in the code
- $\vec{\chi}_t$ is a vector of controls

Effect of Commits on User Activity

	(1) activitylogcount	(2) eventlogcount
fileschanged	-5530.7	-60589.4*
percentile	(-1.91)	(-2.26)
insertions	4868.8*	47053.7*
percentile	(2.12)	(2.22)
deletions	2778.9	29970.6
percentile	(1.29)	(1.50)
weekend	-14708.0***	-79769.8***
	(-16.48)	(-9.65)
_cons	22396.6***	224482.8***
	(25.05)	(27.12)
N	474	475

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Effect of Commits on User Activity

	(1) avguseractivity	(2) avguserevents
fileschanged	-10.40*	-148.6
percentile	(-2.01)	(-1.52)
insertions	2.817	79.92
percentile	(0.69)	(1.03)
deletions	4.670	49.47
percentile	(1.21)	(0.68)
weekend	-0.297	257.0***
	(-0.19)	(8.49)
_cons	25.00***	340.6***
	(15.58)	(11.24)
N	474	475

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Specification

$$y_{tx}^{i} = c_0 + \gamma M_t + \beta weekend_t + \epsilon_t \quad (2)$$

- Lag variable $x \in [1, 30]$.
- Examines how commits on day t affect user behavior in time $t + x$.

Average Activity Logs Per Distinct User

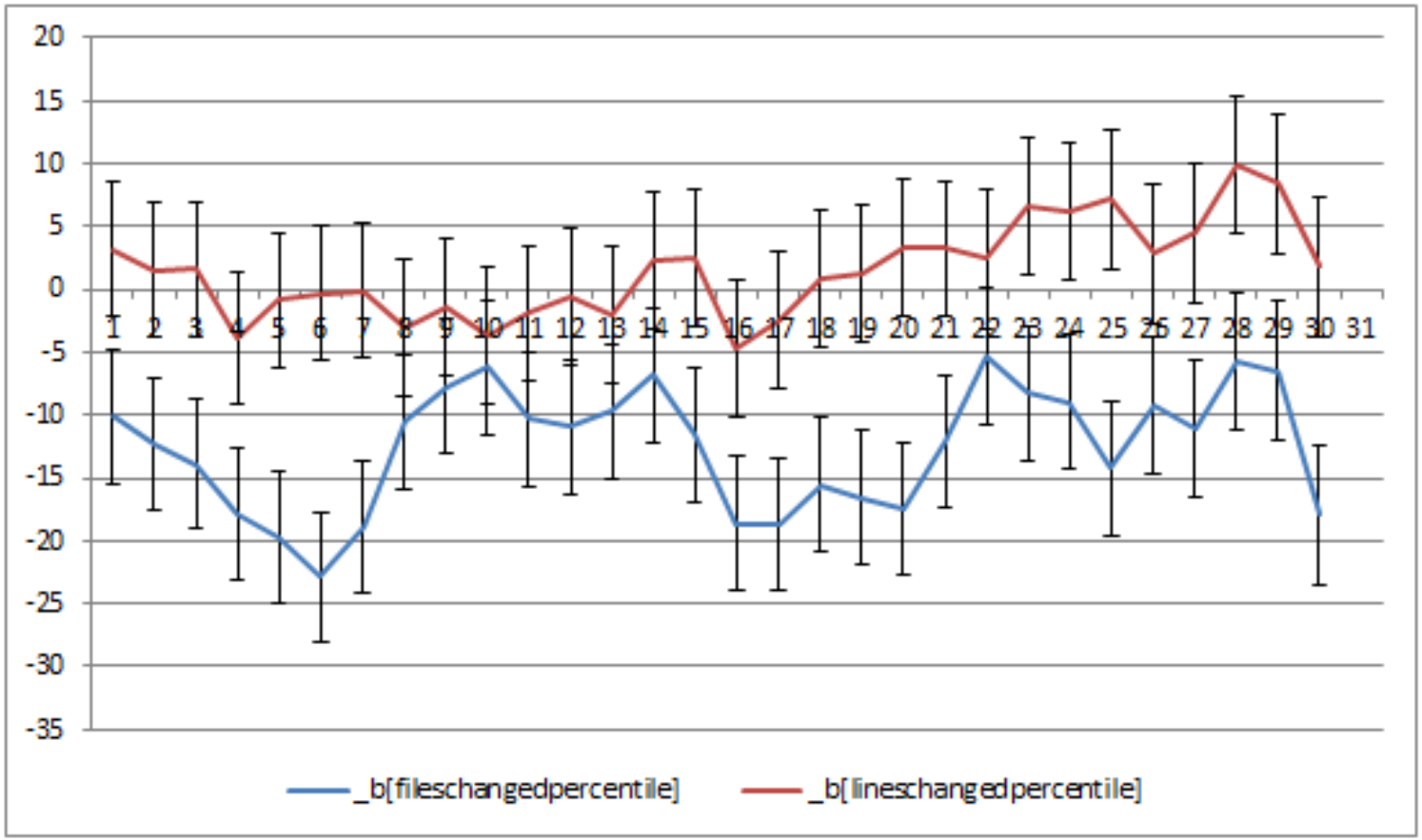


Figure: γ Coefficients with Varying Lags, Regressed on Average Activity Logs per User

Average Event Logs Per Distinct User

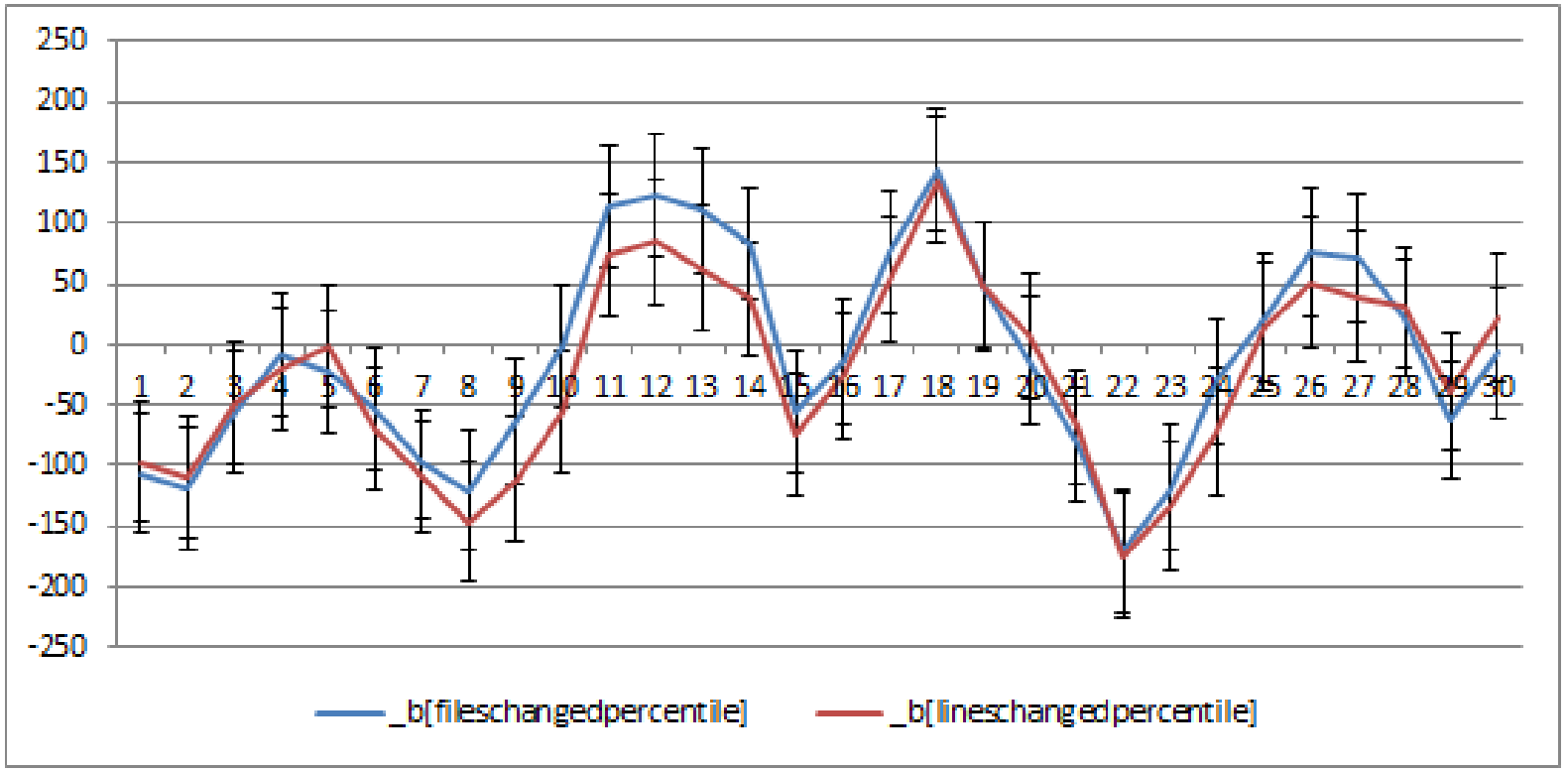


Figure: γ Coefficients with Varying Lags, Regressed on Average Event Logs per User

Insertions and Deletions

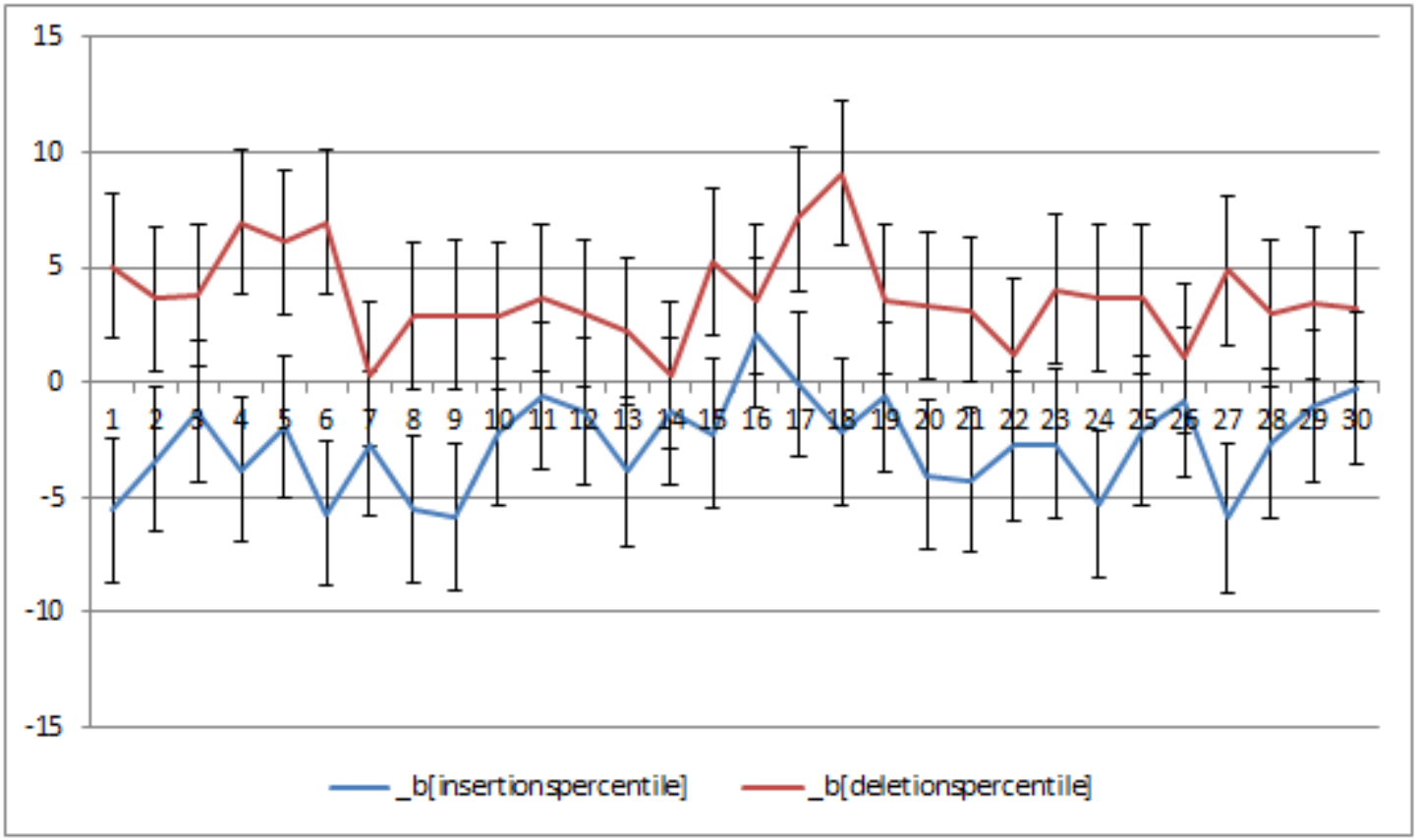


Figure: γ_0 and γ_1 Coefficients with Varying Lags. Regressed on Average Activity Logs per User

Problems with Macro-Level Data

- Nothing more than correlations
- Possibly complex, unknown mechanisms for how results come about

The Case for Micro-Level Data

- Panjiva has data on the page and time that any action was performed.
- Code changes (commits) can be thought of as exogeneous shocks.
- Almost all changes are unannounced
- Only extremely large changes are announced on blog (less than 1% of Panjiva's total pageviews come from blog).

Search Controller

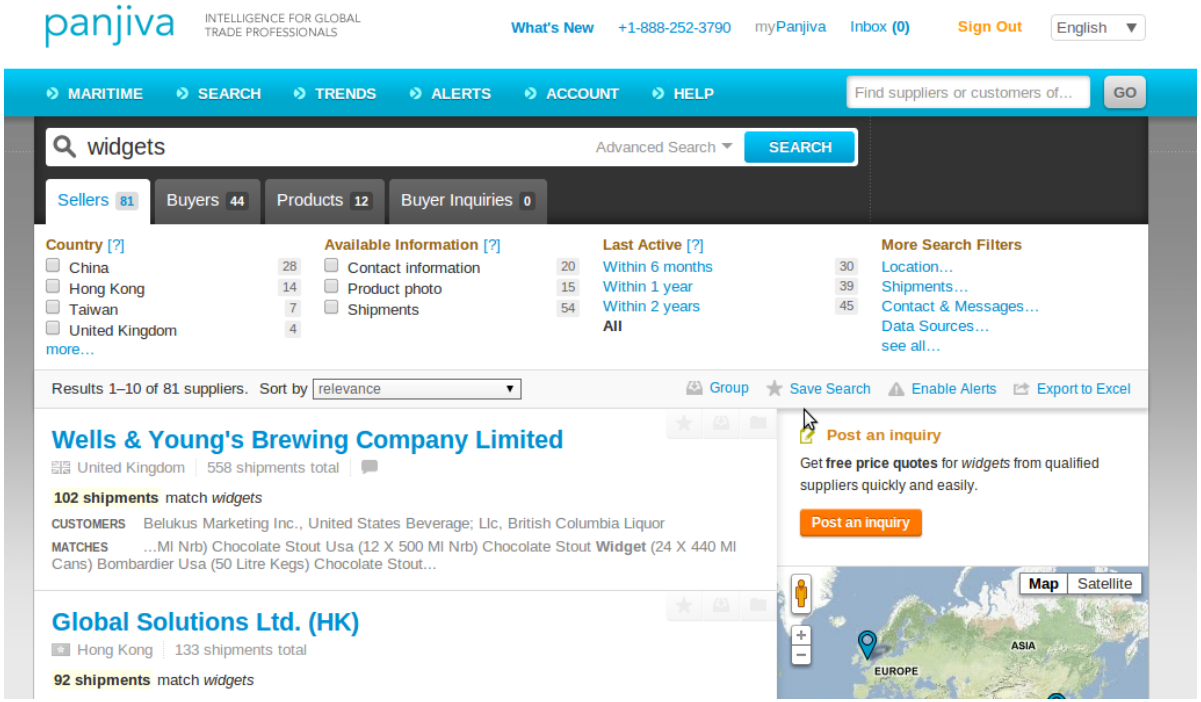


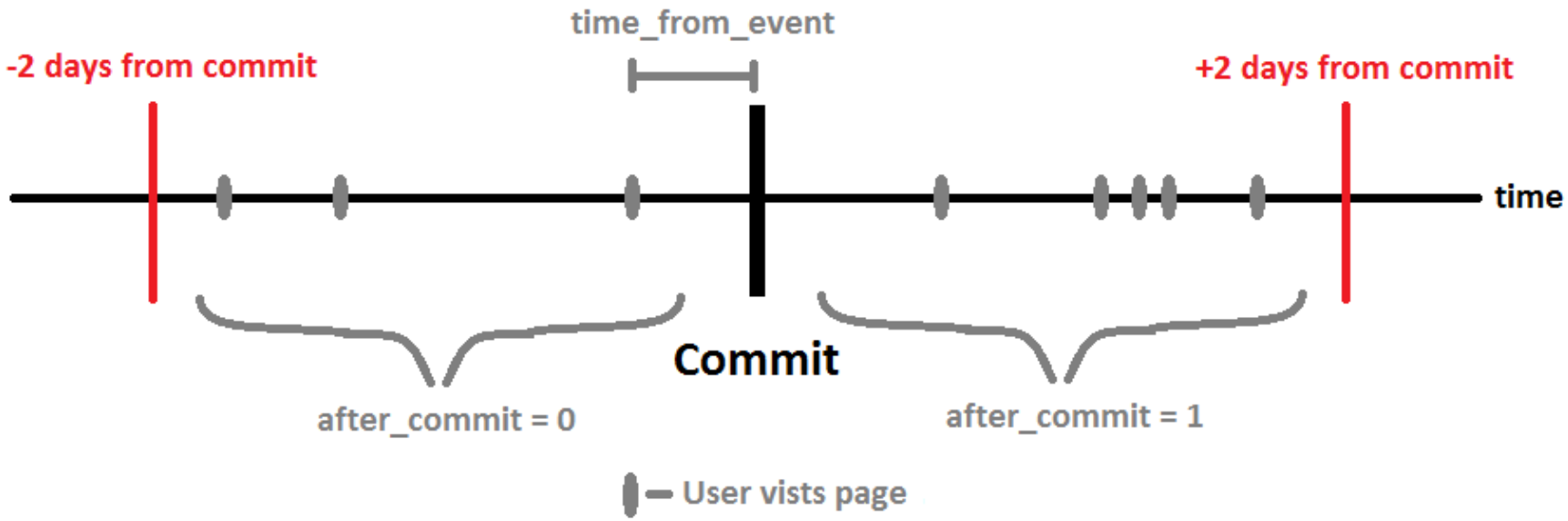
Figure: The Search Page, Panjiva's most trafficked page, provides functionality for finding suppliers, buyers, products, and buyer inquiries.

Specification

$$num\_views\_day\_after_{it} = c_0 + \beta_0 \mu_{it} + \beta_1 hour\_dummies_{it} + \epsilon_{it} \quad (3)$$

- i indexes the ith commit and t is the tth view in 4 day window surround the ith commit.
- μ_{it} is either *after_commit* or *time_from_event*.

Visualization of Variables



Search Regression Results

	Dependent Variable: num_views_day_later			
	(1)	(2)	(3)	(4)
after_commit	5.541***		3.378***	
	(17.91)		(11.20)	
time_from_event		0.0000737***		0.000110***
		(43.76)		(66.60)
created_at_hour dummies?	No	No	Yes	Yes
_cons	156.9***	159.9***	138.3***	139.7***
	(731.32)	(1033.47)	(127.71)	(130.50)
N	2345617	2345617	2345617	2345617

t statistics in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Specification

$$
\begin{aligned}
num\_views\_day\_later_{it} &= \beta_0 after\_commit_{it} + \beta_1 insertionspercentile_{it} + \\
&+ \beta_3 insertionspercentile_{it} * after\_commit_{it} \\
&+ \beta_4 deletionspercentile_{it} * after\_commit_{it} \\
&+ \beta_5 hour\_dummies_{it} + c_0 + \epsilon_{it}
\end{aligned}
$$

- Interaction term coefficients β_3 and β_4 give incremental percentile coefficients after the commit
- Same specification, but including interaction terms to see additive effect

Differences in Commits

	Dependent Variable: num_views_day_later			
	(1)	(2)	(3)	(4)
after_commit	-10.86***	-13.79***	-325.5***	-305.9***
	(-16.90)	(-22.03)	(-154.63)	(-149.28)
insertionspercentile	-50.97***	-50.76***	-650.6***	-609.5***
	(-54.12)	(-55.34)	(-218.09)	(-209.88)
deletionspercentile	66.31***	63.35***	251.5***	230.8***
	(70.67)	(69.34)	(84.73)	(79.94)
insertionspercentile * after_commit	57.00***	58.05***	393.0***	365.1***
	(41.21)	(43.09)	(89.07)	(85.01)
deletionspercentile * after_commit	-24.16***	-23.64***	2.306	3.295
	(-17.65)	(-17.74)	(0.53)	(0.77)
Hour Dummies?	No	Yes	No	Yes
Controllers Used	Search	Search	All	All
_cons	150.1***	133.2***	606.5***	491.0***
	(336.67)	(116.37)	(420.98)	(130.17)
N	2345617	2345617	3858943	3858943

Effects of Each Controller

$$\begin{aligned} num\_views\_day\_later_{it} &= \beta_0 after\_commit_{it} + \bar{\beta}_1 \overline{controllers}_{it} \\ &+ \bar{\beta}_2 \overline{controllers}_{it}^{T} \times \overline{after\_commit}_{it} \\ &+ \beta_3 hour\_dummies_{it} + \epsilon_{it} \end{aligned}$$

- Looking for the impact of controller k on user activity after a commit
- Want to examine $\Gamma_k = \beta_{2k} + \beta_{1k} + \beta_0 - \beta_{1k} = \beta_{2k} + \beta_0$.
- Standard errors given by:

$$SE_{sum} = \sqrt{SE^2_{cont_{2k}} + SE^2_{ac} + 2Cov(cont_{2k}, ac)}$$

Controller Results

	(1) No Hour Controls	(2) With Hour Controls
Communication	-2.673	5.111
	(-0.50)	(0.98)
My_Panjiva	-41.718*	-54.530***
	(-1.98)	(-2.68)
Profile	-8.167	-6.339
	(-1.30)	(-1.04)
Project	-410.438***	-388.496***
	(-235.55)	(-228.27)
Search	5.541***	1.516
	(4.37)	(1.23)
US_Exports	10.318	-28.946
	(0.28)	(-0.80)
US_Imports	-12.487	-8.73
	(-0.82)	(-0.59)
N	3858943	3858943

Conclusions

- Code changes require particular factors for success
- Users respond negatively to low quality code (total and average user activity decreased when more files were changed)
- Changes to highly trafficked pages have greater chance of success
- Code changes can negatively affect user behavior

Acknowledgements

I would like to thank:

- Jim Psota
- Professor Ellison
- Panjiva, Inc.