Predicting Litigation for Misrepresentation and Omission in IPO Filings

Prepared By: James Burden, Matthew Dunne, Jose Nueno, Hairun Wang

Background

Lyft accused of misleading investors, inflating IPO share price

Alibaba Agrees to Settle Its IPO Lawsuit

Alibaba to pay $75 million for the settlement

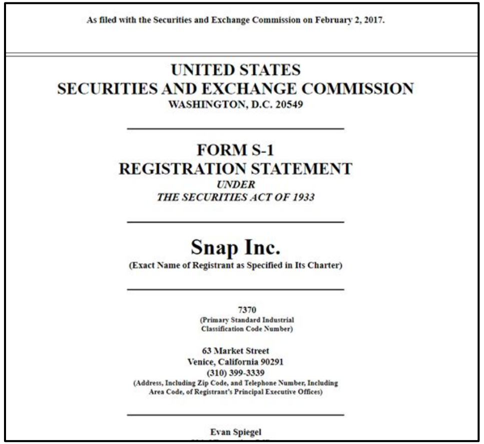
Snap Shareholders Can Pursue Claims That IPO Hid Crucial Information

Facebook to Pay $35 Million to Settle Lawsuit Over IPO

GRAHAM SCHOOL UCHICAGO

What are their IPOs getting sued for?

For statements made in their IPO registration statement a.k.a. S-1 filing.


As filed with the Securities and Exchange Commission on February 2, 2017.

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
WASHINGTON, D.C. 20549

FORM S-1
REGISTRATION STATEMENT
UNDER
THE SECURITIES ACT OF 1933

Snap Inc.
(Exact Name of Registrant as Specified in Its Charter)

7370
(Primary Standard Industrial
Classification Code Number)

63 Market Street
Venice, California 90291
(310) 399-3339
(Address, Including Zip Code, and Telephone Number, Including
Area Code, of Registrant's Principal Executive Offices)

Evan Spiegel

- Business operations
- Financial condition
- Results of operations
- Risk factors
- Management

These filings must clearly describe all important information to inform potential investors pre-IPO

Public companies can be **sued for information misrepresentation and/or omissions** in these filings, even if they were unintentional.

Alibaba: "...**failed to provide full public disclosure** of its securities…"

Groupon: "...a shareholder who **accused the company of misleading investors**…"

Facebook: "...**prospectus contained untrue statements of material facts**…"

GRAHAM SCHOOL UCHICAGO

Motivation for Research

Problem Statement: Can we ...

Research Purpose: We want to ...

Predict whether companies will get sued based on the language used in IPO filings.

+

Identify what particular NLP metrics can be improved to avoid getting sued

=

Proof of concept that NLP variables alone can predict getting sued.

GRAHAM SCHOOL UCHICAGO

Our Data: 205 sued vs 798 not sued; 48 predictors

Parsing and Merging → **Filtering** → **Data cleaning**

Parsing and Merging

IPO Financial Data
Source format: CSV
Provides non-NLP predictors

+

Lawsuit Statement
Source format: XML
Provides response label & manually coded

+

Registration Statement
Source format: XML
Used to derive NLP predictors

Filtering

1. **Federal court** only
2. Must have a **drop in stock price below its IPO within first 3 years**
3. IPOs between **1995-2015**
4. **S-1 filing** only

Data cleaning

Removed anomalies and **incomplete data** (especially less than 3 years worth of price data)

Unit of Analysis: the IPO (one per company)
Dependent Variable: Sued vs. Not Sued (1 or 0)

Hypothesis: subsequent share price is all that matters

This is surprisingly not the case, so there seems to be some scope for the impact of NLP variables.

Largest price drop in % bins	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Not sued	0.04	0.04	0.06	0.09	0.07	0.07	0.12	0.13	0.15	0.23
Sued	0.02	0.02	0.05	0.07	0.08	0.09	0.08	0.15	0.18	0.28

This suggests to us that although price drops are a necessary condition to file a lawsuit under the applicable sections of the law (as required evidence of economic loss), they are by themselves not a sufficient condition.

Parsing S-1 filings for NLP metrics - 3 categories:

Degree of Disclosure

- Average sentence length
- Paragraph count
- Sentence count
- Word count

Readability

- Gunning-Fog index

Sentiment

- Negative word count
- Positive word count
- Overall Sentiment*
- Constraining word count
- Difficult word count
- Interesting word count
- Litigious word count
- Modal word count

Metrics based on previous literature

Specifically within field of finance

GRAHAM SCHOOL UCHICAGO

*Overall sentiment = (positive word count – negative word count) / (total word count).
It is then normalized from 0 to 1 with 0 being the most negative and 1 being the most positive.

A snapshot of the data frame:

	Ticker	Name	Sued	IndustrySector	MarketCapatOff	largest_price_percent_DROP	avg_sentence_len	con	difficult_words	inte	lit	mod	pos	neg
0	A	AGILENT TECHNOLOGIES INC	1	Industrial	13110.00	37	30	629	23423	203	1381	891	1130	2068
1	ABG	ASBURY AUTOMOTIVE GROUI	0	Consumer, Cyclical	561.00	65	33	474	10873	70	525	451	577	972
2	ACLA	ACLARA BIOSCIENCES INC	1	Consumer, Non-cyc	654.63	85	31	334	13240	109	797	549	758	1495
3	ACLS	AXCELIS TECHNOLOGIES INC	0	Technology	2135.10	85	31	442	13617	95	690	638	729	1405
4	ACME	ACME COMMUNICATIONS IN	0	Communications	385.25	83	34	716	16840	210	878	689	642	1461
5	ADBL	AUDIBLE INC	1	Communications	225.03	96	26	307	9084	98	451	278	400	800
6	ADLR	ADOLOR CORP	0	Consumer, Non-cyc	402.74	46	33	463	12260	137	694	653	709	1388
7	ADNC	AUDIENCE INC	0	Technology	329.68	83	36	846	20007	229	1062	1050	1204	2436
8	ADRO	ADURO BIOTECH, INC.	0	Consumer, Non-cyc	1020.01	86	41	1166	27067	471	1930	1615	1615	3538
9	ADSC	ATLANTIC DATA SERVICES INC	0	Technology	165.99	89	30	149	6003	69	336	145	216	483
10	ADUS	Addus HomeCare Corp	1	Consumer, Non-cyc	104.96	26	33	863	23745	232	1252	899	1169	2613

However, NLP metrics do little to separate sued vs non-sued

Sued score - Non-Sued score for each NLP metric:

word count		sentence count		paragraph count		avg sentence length		difficult %		litigious %		negative %	
mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
-10,839	-20,424	-247	-694	26	-58	-1.35	0.28	1.31	0.51	0.1	0.5	0.1	0.2

positive %		constraining %		uncertain %		readability		sup %		interesting %		modal %	
mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
0.1	0.0	-0.03	-0.05	0.0	0.1	-0.01	0.13	0.0	0.0	0.0	0.0	0.0	0.0

GRAHAM SCHOOL UCHICAGO

We created a matched data set to better isolate the effects of NLP metrics on litigation: 170 sued vs 220 non-sued

Same **industry**, same **market capitalization decile**, and whose **largest price drop** is within 25% of sued companies.

Sued company #1	Non-sued company **A**	Non-sued company B	Non-sued company C
Sued company #2	Non-sued company **A**	Non-sued company E	N/A
Sued company #3	Non-sued company D	N/A	N/A
~~Sued company #4~~	N/A	N/A	N/A

Initial promise on predictive models using NLP metrics alone

	Sued		Non-sued	
	Recall	Precision	Recall	Precision
Logistic Regression (47% threshold)	0.67	0.52	0.52	0.67
Random Forest	0.59	0.59	0.69	0.69
AdaBoost	0.48	0.45	0.60	0.63
Support Vector Machine	0.43	0.58	0.76	0.64

GRAHAM SCHOOL UCHICAGO

Adding non-NLP variables does not improve performance

	Sued		Non-sued	
	Recall	Precision	Recall	Precision
Logistic Regression (43% threshold)	0.84	0.43	0.13	0.53
Random Forest	0.69	0.59	0.68	0.77
Naïve Bayes	0.68	0.59	0.50	0.60
AdaBoost	0.56	0.59	0.67	0.64
Support Vector Machine	0.51	0.60	0.75	0.67
Logistic Regression	0.29	0.43	0.70	0.57
Logistic Regression w/ Regularization	0.31	0.48	0.75	0.59

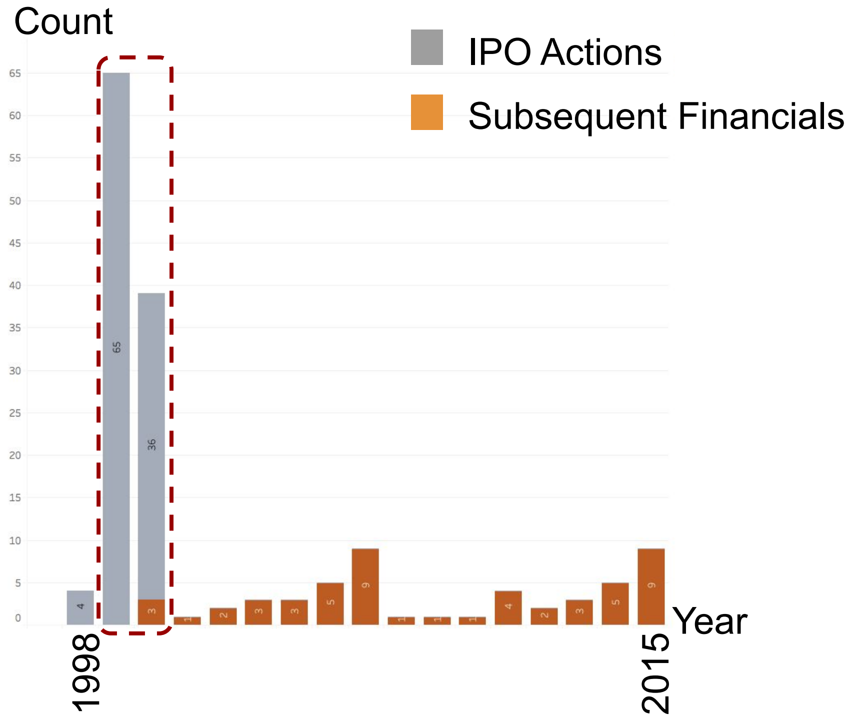
However, NLP-only model does not extrapolate well to full dataset

	Sued		Non-sued	
	Recall	Precision	Recall	Precision
Logistic Regression (47% threshold)	0.22	0.19	0.78	0.81
Random Forest	0.04	0.14	0.94	0.81

We therefore conclude that NLP metrics alone have limited predictive power and has failed as a proof of concept.

Topic modeling identifies 2 kinds of lawsuits based on lawsuit content alone - dotcom bubble and others.

Inferred Topic	Sample Key Words	Number of lawsuits
Actions at IPO	Issuer, underwriter, commission, compensation, price	105
Subsequent Financial Statements	Result, quarter, fact, financial, time, revenue, false, report	53
Procedural Language	Case, action, party, state, serve, file, date, service	16



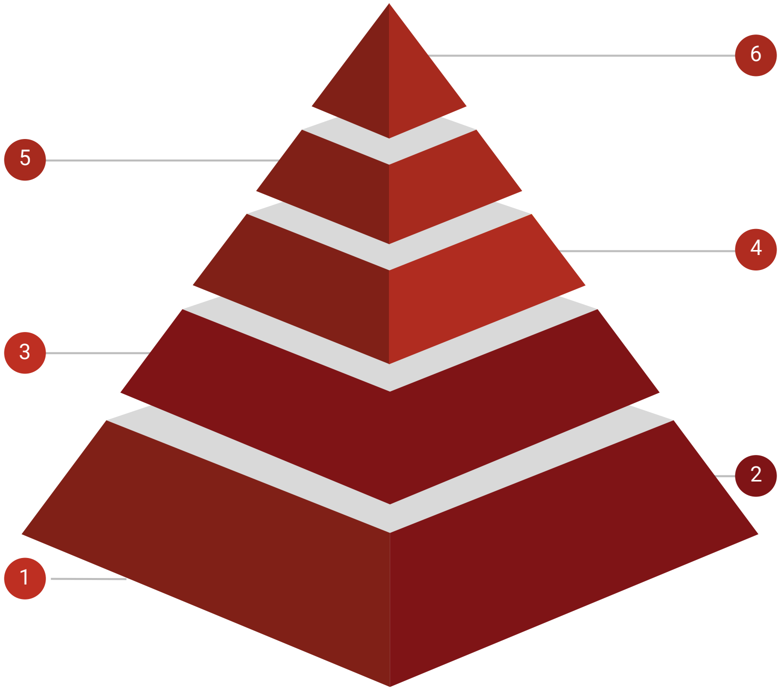
We tried multiple approaches in the project

1 — 2 — 3 — 4

1. Combining different data sources
2. Parsing registration statements
3. Creating a matched dataset to isolate the effects of language on litigation
4. Various combinations of models and predictors on matched and full data set

GRAHAM SCHOOL UCHICAGO

But challenges are inevitable



5 There isn't much signal in the text. Even a human being couldn't predict the language that Lyft would be sued on

3 Certain non-NLP predictors like price movement you don't know in advance

1 Not a lot of data, which also limits the number of predictors used in a model

6 How do you detect or test for omission?

4 Market conditions change and have changed over the last 20 years **(topic modelling)**

2 Registration statements are very long and extremely similar

GRAHAM SCHOOL UCHICAGO

Findings

EDA
Little separation in price drop, NLP metrics in sued companies

Topic Modeling
We really have two populations (dot-com bubble and other) and those market conditions are probably the drivers of litigation

Methodology and Models
Some initial promise in predictive models (for just NLP), but does not translate to full data

Methodology and Models (Cont.)
Even when we incorporate subsequent information you'd think is important (price drop) it doesn't help much

Insights
NLP really doesn't matter or at least doesn't have much predictive value

Here are recommendations for future work

01

While other NLP tools could be employed, such as native text analytics (count vectorizer), we do not believe this would be a worthwhile effort given what we have seen thus far.

02

Research should instead be focused on price-related metrics, analyst forecasts, market conditions, etc.

GRAHAM SCHOOL UCHICAGO

“I **can calculate the motion** of **heavenly bodies**, **but not the madness** of **people**.”

~ Isaac Newton

GRAHAM SCHOOL UCHICAGO

https://www.businessinsider.com/isaac-newton-lost-a-fortune-on-englands-hottest-stock-2016-1

Appendix - Word Lists to Capture Sentiment

Fin-Neg: There are in total 1,202 words, including *litigation, discontinued, unpaid, investigation, misstatement, misconduct, forfeiture, serious, allegedly, deterioration,* and *felony*.

Fin-Pos: There are in total 264 words, including *achieve, attain, efficient, improve, profitable,* or *upturn.* Importantly, to account for simple negation we must also parse for the words *no, not, none, neither, never, nobody* that immediately precede positive words.

Fin-Unc: There are in total 123 words, including *approximate, contingency, depend, fluctuate, indefinite, uncertain,* and *variability.*

Fin-Lit: There are in total 95 words, including *claimant, deposition, interlocutory, testimony, tort*, *legislation* and *regulation.*

MW-Strong: There are in total 19 words, including *will, always, best, clearly, definitely, highest, lowest, must, strongly, uncompromising, undisputed, unsurpassed*

MW-Moderate: There are in total 12 words, including *can, frequently, generally, likely, often, probably, rarely, regularly, should, tends*

MW-Weak: There are in total 27 words including *almost, apparently, appears, could, depend, may maybe, might, nearly, occasionally, perhaps, possibly, seldom, sometimes, somewhat, suggest, uncertain*

Constraining: There are in total 184 words, including *abide, commitment, comply, constrained, dependent, impose, limiting, mandatory, necessitate, obligated, prohibit, requirement, unavailable*