Topic Modeling Consumer Complaints for Risk Analysis

Government Advances in Statistical Programming (GASP) Workshop

B.J. Bloom | *September 23, 2019*

Goal: use consumer complaints to inform risk analysis for Federal Reserve Board (FRB)

- Problem: what is the best way to identify emerging risks in consumer financial products?
  - Consumer complaints may shed light on consumer experience in a unique way
- CFPB complaint volume (over 1 million complaints since 2011) is far higher than FRB complaint volume (20,000 since 2012)
  - Our data sharing agreement gives us access to the un-redacted consumer complaint narratives submitted to the CFPB
  - The larger complaint database (CFPB) is more conducive to statistical analysis and broader trend identification

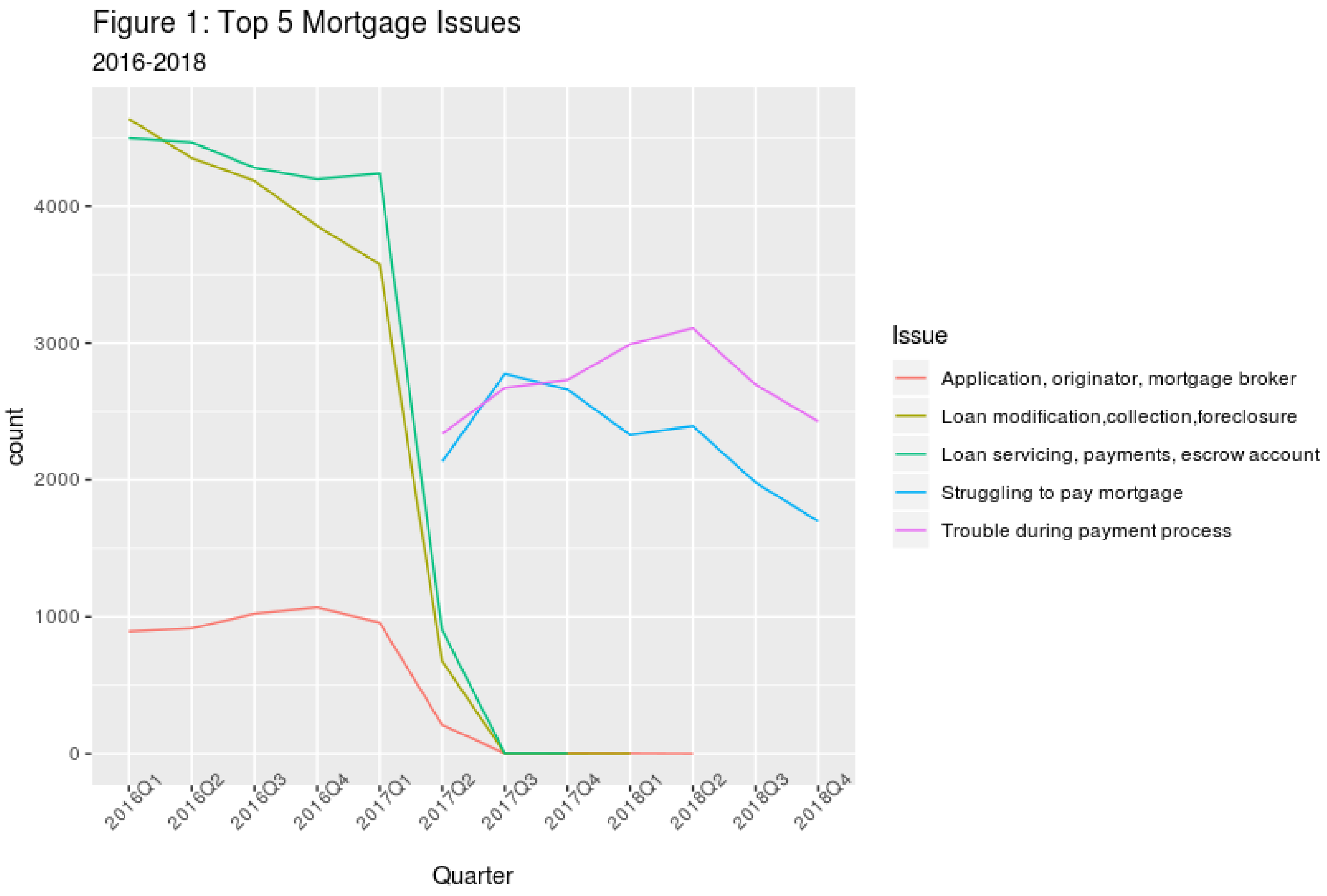
Consumer complaints have some inherent limitations as a data source

- Limitation to using consumer complaint data for risk analysis
  - Complaints are not a representative sample of consumer experience
  - Complaints vary in salience
  - Complaints do not necessarily mean a company did something wrong
- In risk analysis, we assess meaningful trends, not the veracity of each complaint
- Complaints are only one component of overall risk analysis of these product

The CFPB collects a lot of metadata related to complaints but there are also some limitations

- Skewed distribution of complaint categories
  - Some categories too broad, some too narrow
  - Very few “Goldilocks” categories
- Some duplicate or ambiguous categories
- Way consumers talk about financial products differs from how regulators think about them
- CFPB re-categorized their metadata in April 2017

Time series of mortgage issues shows one issue with relying on the CFPB metadata



Topic modeling: uncovering latent topics within corpus of complaints

- Solution to limitations of metadata: topic modeling
- Topic modeling is a type of Natural Language Processing (NLP) algorithm which assumes that, for a given set of documents (corpus), these documents contain a set of topics, each of which are composed by a set of words
- The particular topic modeling technique used is **Latent Dirichlet Allocation (LDA)**, a Bayesian model which, given K number of topics, will iteratively assign words to topics and topics to documents as it updates information contained within the documents.

The model is complicated but involves inferring from an assumed generative model

- Known parameters (or priors)
  - α: Prior assumption of topic distribution of documents
  - β: Prior assumption about the word distribution of each topic
- Unknown (latent) parameters
  - K: number of topics
  - θ: Topic distribution in each document
  - ϕ: Word distribution of each topic
  - z: Word topic assignment

$$p(\theta, \phi, z | w, \alpha, \beta) = \frac{p(\theta, \phi, z | \alpha, \beta)}{p(w | \alpha, \beta)}$$

The basic conceptual process

Documents
(contain words)

Topic Model
Algorithm

LDA fits “best”
topics

Topics (most probable words)

t.1	Account, close, request
t.6	Debt, collect, owe, notice
t.14	Payment, late, due

Topic assignments (per each document)

t.1	0.71
t.6	0.09
t.14	0.20

Topic model conceptual example: newspaper articles with no headlines

- Imagine you had 5 years' of newspaper articles from the Washington Post with no headlines
- If you create a topic model with 3 topics, output could be:
  - Topic 1: election, poll, moderate, healthcare, rules
  - Topic 2: goal, score, team, fans, rules
  - Topic 3: theater, play, drama, fans, review
- This allows you to identify topics, review articles associated with certain topics, and look at trends in topics over time
  - However, it doesn't replace an actual headline of an article

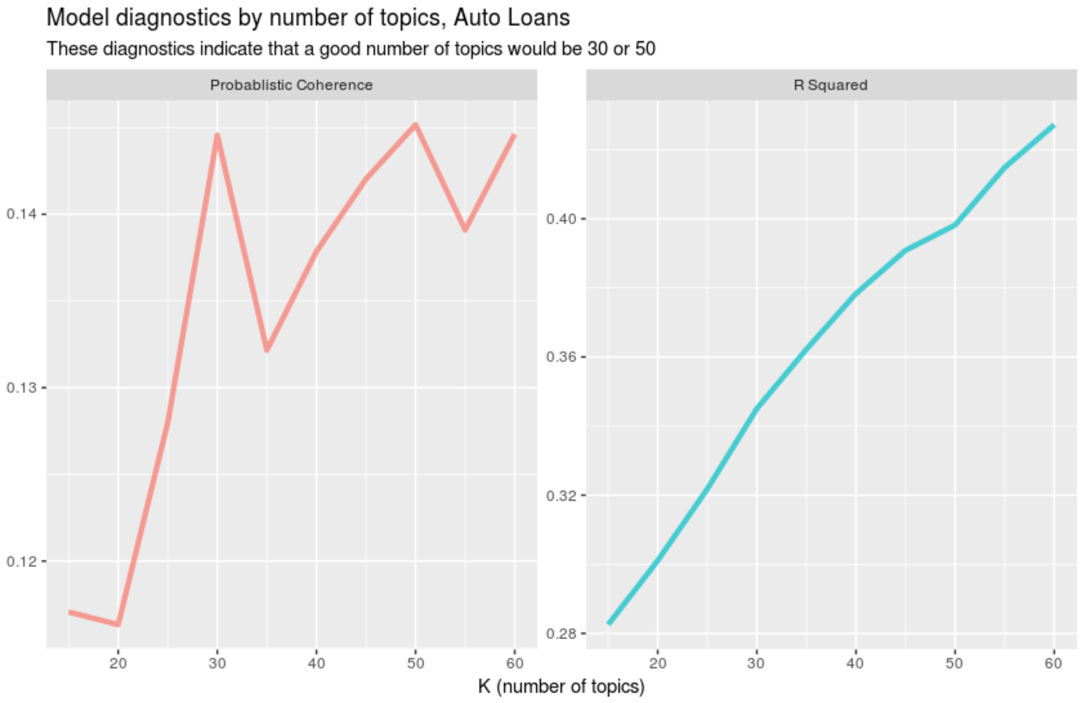
{textmineR} package in R is used for these topic models due to intuitive model diagnostics

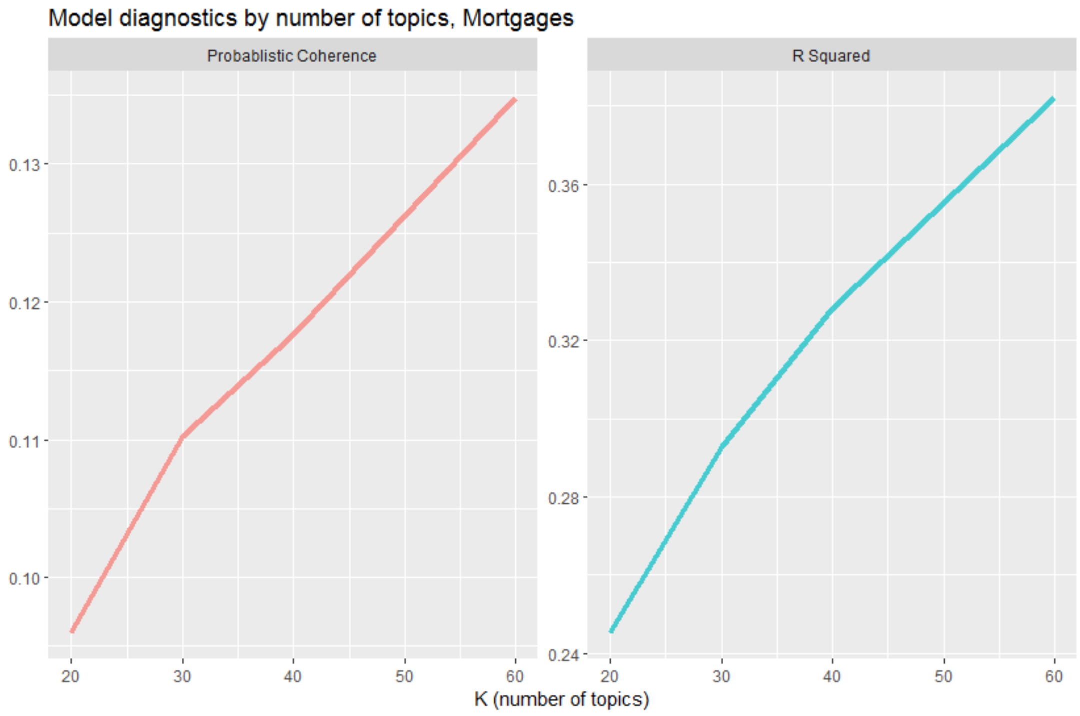
- Biggest challenges in topic modeling are 1) understanding quality of the model and 2) selecting the optimal number of topics
- Main general NLP packages in R: tm, tidytext, quanteda, udpipes, textreuse, text2vec, SnowballC, textrank
- Main topic modeling (LDA) packages in R: lda, topicmodels, **stm**, **textmineR**
  - **stm** can incorporate metadata into the topic probabilities (the prior, usually)
  - **textmineR** includes two diagnostic measures (R-squared and probabilistic coherence) that help assess the model

Intuitive explanation of probabilistic coherence (for selecting “best” model)

- Probabilistic coherence measures how associated words are in a topic, controlling for statistical independence
- For each pair of words ***{a,b}*** in the top M words in a topic, probabilistic coherence calculates $P(b|a)-P(b)$, where ***{a}*** is more probable than ***{b}*** in the topic.
- Probabilistic coherence measure averages this calculation across M number of words in a topic
- Selection of K (number of topics) involves **fitting multiple topic models** and finding the optimal average probabilistic coherence measure (across all topics) as well as optimal R-squared value

These diagnostics work to a varying degree depending on the model/product

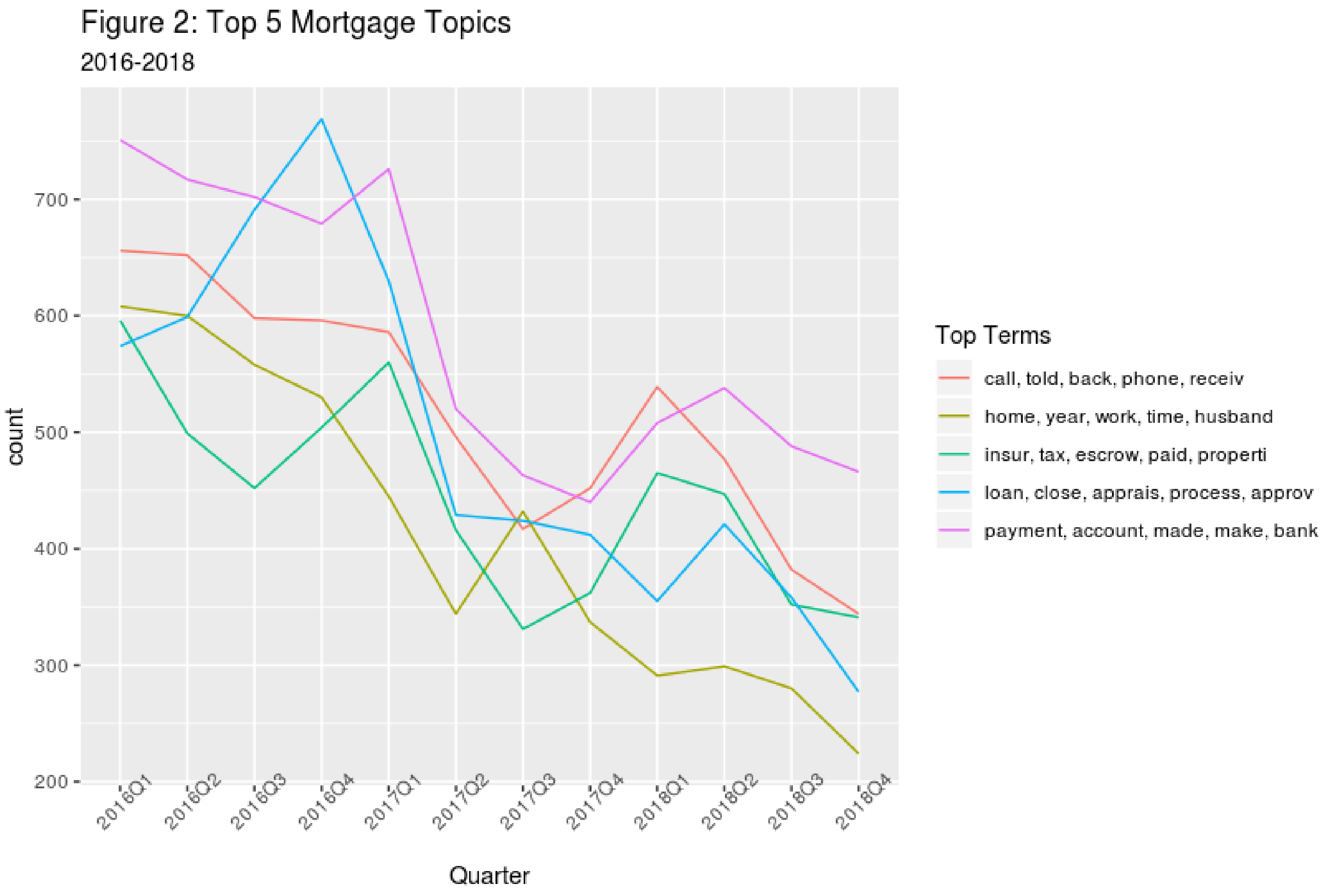




Some practical business concerns took precedence over the “optimal” technical solution

- Built 10 separate models (one for each product), but K was limited to at most 40 topics
  - Hard to explain, summarize, or create trend lines for more than that
- Topic models output a distribution of multiple topics, but each complaint was hard-coded with one topic (highest probability)
  - Avoids double-counting
  - Highlights actual trends
- Some manual review to get better understanding of top 5 topics

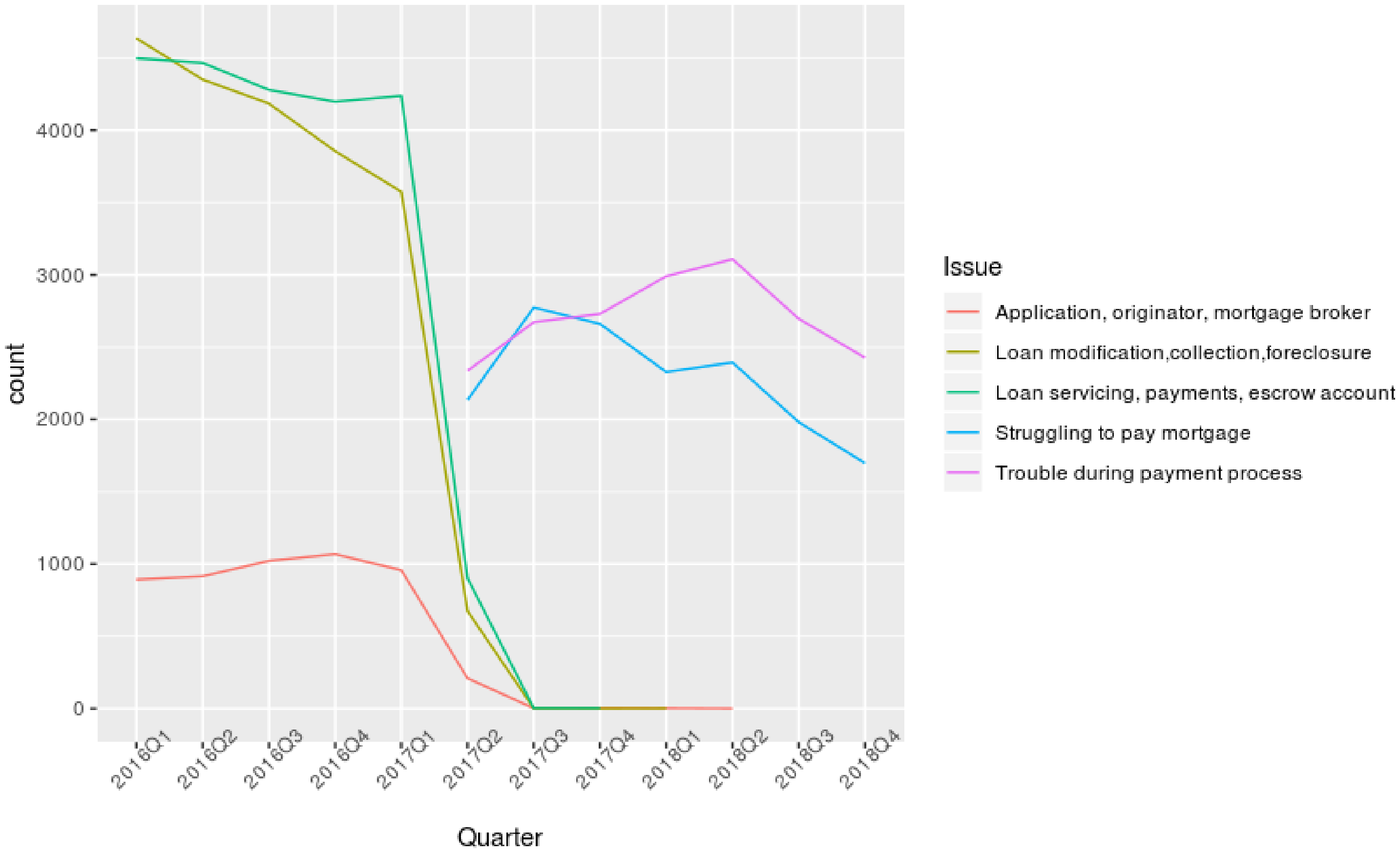
Topic model output helps smooth out trend lines for when the CFPB categories changed



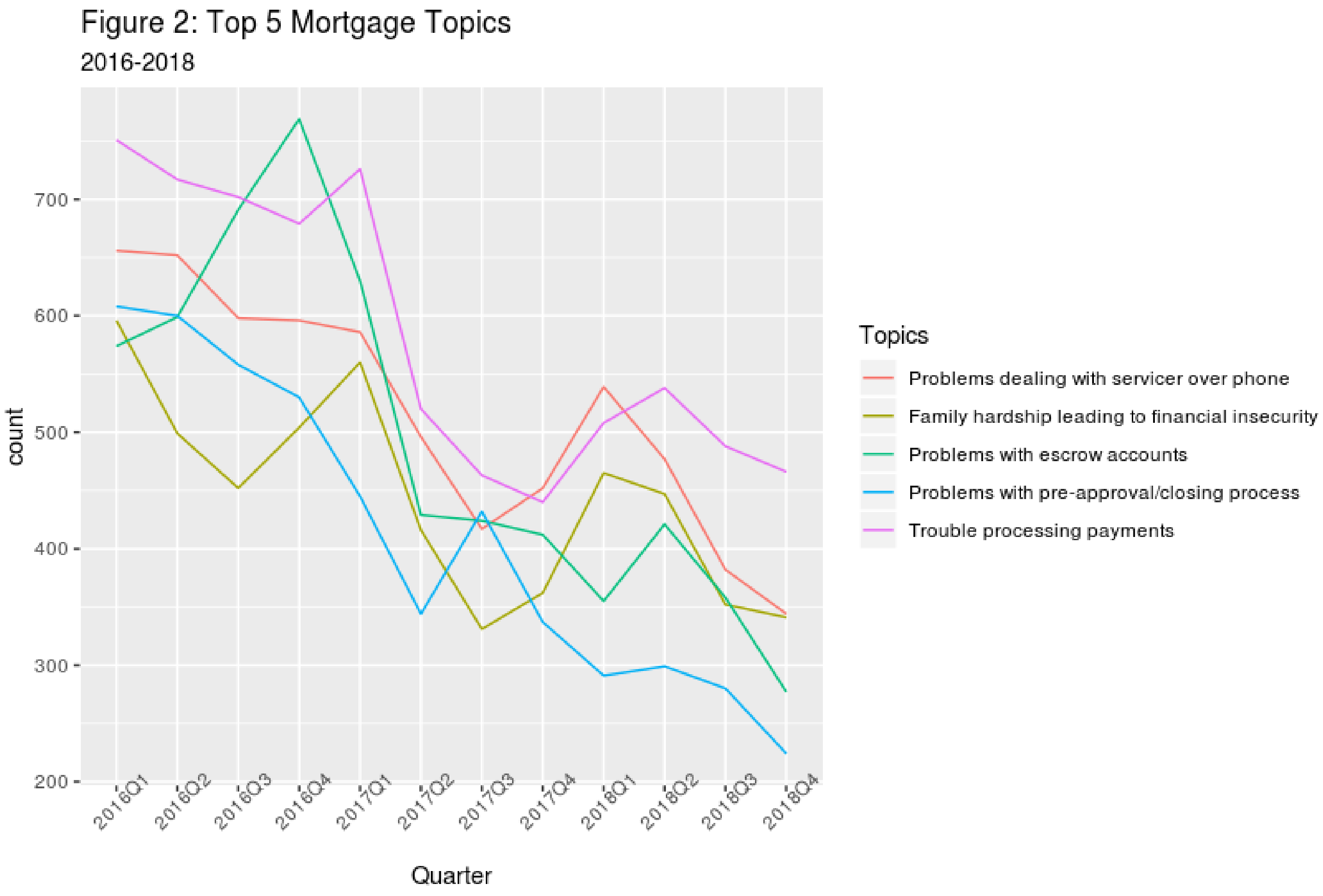
Comparison to previous figure

Figure 1: Top 5 Mortgage Issues

2016-2018



We can come up with more descriptive phrases for topics through some manual review



Results of models incorporated into broader risk report for key stakeholders and R Shiny dashboard

- Our team (Risk & Surveillance) compiles a set of risk reports; Consumer Complaints Report is one section
  - Uses both FRB and CFPB complaint data for broad market overview of consumer complaint risk landscape
- Purpose is to understand emerging risks based on data from multiple sources
- R Shiny dashboard allows end users to explore complaint topics in more depth

Advantages of using topic modeling on consumer complaint data

- A good way of identifying **emerging issues** someone may not thought of ahead of time
  - Topic modeling can be effective way of overcoming limitations of metadata
- Categories are consistent over time
- Purpose is to look at broad changes in product markets to inform key stakeholders of potential risks
  - Best data we have to capture consumer experience by product
  - Good initial indicator for risks of potential consumer harm
  - Real-time data that identify trends sooner that other data collection methods

Contact information

B.J. Bloom

Federal Reserve Board

benjamin.j.bloom@frb.gov

@bj_bloom (personal account)

github.com/bjbloom (personal account)