Noname manuscript No.
(will be inserted by the editor)

A Tale of Four Cities

Exploring the Soul of State College, Detroit, Milledgeville and Biloxi

Karsten Maurer · Dave Osthus · Adam Loy

the date of receipt and acceptance should be inserted later

Abstract Can data help us explore and expose the soul of the community? This was the challenge posed by the 2013 Data Exposition. The Knight Foundation, in cooperation with Gallup, furnished data from 43,000 people over three years (2008–2010) in 26 communities, which we explored in an effort to discover variables associated with community attachment. Our analysis focused on four cities that stood out after out initial exploration of the data set: State College, PA, Detroit, MI, Milledgeville, GA and Biloxi, MS. We present our use of survey-weighted binned scatterplots to graphically explore the association between an individual's community attachment and perceived economic outlook. Additionally, we present a few other analyses we found interesting during our initial exploration which we view as a collection of "short stories."

Keywords 2013 ASA data exposition · Exploratory data analysis · Survey data · Reproducible research · Statistical graphics · Binned scatterplots

1 Introduction

Can data help us explore and expose the soul of a community? This was the challenge the Joint Statistical Meetings' 2013 Data Expo posed to teams of statisticians. The Knight Foundation provided survey data on approximately 43,000 respondents from 26 communities between 2008 and 2010. Our team first cleaned and organized the data, then identified communities with interesting stories, as told through statistical graphics. Four cities stood out: State College, PA, Detroit, MI, Milledgeville, GA and Biloxi, MS. The years the survey was administered coincided with the deepest point of the U.S. financial crisis. We presumed the markings of the U.S. financial crisis could be seen in these data. This presumption led us to focus our data exploration on the relationship between individuals' economic outlook and community attachment levels. Our primary narrative is the relationship between economic outlook and community attachment for the four cities previously mentioned. Not unlike a collection of short stories, our secondary narrative is a motley assortment of discoveries found in these data.

2 Data

2.1 What data were provided?

The Knight Foundation, in cooperation with Gallup, collected data from 43,000 people over three years (2008–2010) in 26 communities across the United States (Hofmann & Wickham, 20XX). Each year, a sample of individuals from each of the 26 communities was selected and administered the Soul of

K. Maurer · D. Osthus
Iowa State University, Ames, IA, USA

A. Loy
Lawrence University, Appleton, WI, USA
E-mail: adam.m.loy@lawrence.edu

the Community Survey (SoCS). Survey items in the SoCS were related to demographic information, geographic information and ratings of the community with respect to an individual's loyalty, education, civic involvement, aesthetics and economic outlook.

Between 398 and 1,589 individuals were administered the SoCS for each community/year, with the median number of individuals equal to 401. The SoCS has almost 200 survey items. Thus, at the individual level, this is a "rich" data set, with a substantial breadth of information collected on nearly 43,000 individuals. Though there is a large breadth of information at the individual (or within community/year) level, there is less information at the community level. There is, however, more structure. This is because these data can be considered longitudinal at the community level, but not at the individual level. Thus, we can track the evolution of communities over time, but not individuals. The SoCS also provides temporal and spatial information. The SoCS was administered for three years, from 2008 to 2010. Though this provides an opportunity to track how information evolves over these three years, this cannot be considered a temporally "rich" data set. As there were 26 communities administered the SoCS, spatial information is also available to us. The 26 communities provide coverage for the continental U.S., but 26 spatial locations cannot be considered a spatially "rich" data set. For these reasons, we did not consider fitting temporal or spatial models.

Two different survey weights were provided in the SoCS: the survey weight (WEIGHT) and the projected survey weight (PROJWT). Each weight was created to account for age, sex, race, ethnicity and education, and to account for nonresponse and non-coverage. When used, the survey weights will create unbiased, representative results. WEIGHT should be used when analyzing data from one community or across two or more communities. PROJWT should be used when analyzing results for a combination of aggregation of two or more communities as PROJWT places communities in the correct proportions to one another.

2.2 Organizing the data sets

The data were provided in three comma separated value files; one for each year. Additionally, two codebooks were provided—one for 2008–2009 and one for 2010. The codebooks proved useful, as names of common survey items were not always consistent across years. For example, the variable "Community Attachment" was coded as "CCE" in 2008, "CA" in 2009 and "CCA" in 2010. We created a common and intuitive naming convention of survey items, allowing us to merge the three data files. The names we attached to a subset of survey items can be found in the "Variable Names" column of Table 1. The "Variable Names" column is especially important for any reader following along with the corresponding R code used to process and analyze these data.

After a common naming scheme was assigned, different formats of these data were considered. When exploring these data at the individual level, we worked with a data set where each row represented an individual. This data set had 42,941 rows.

As mentioned earlier, these data can be considered longitudinal at the community level. Thus, we also organized our data so that each row represented a community/year, resulting in a data set with 78 (26×3) rows. Individual-level survey items were aggregated to the community/year level by calculating survey weighted means, using the WEIGHT variable. More specifically, let w_{ijt} be the WEIGHT variable for individual $i = 1, \ldots, N_{jt}$ from community $j = 1, \ldots, 26$ for year $t = 2008, 2009, 2010$. Let X_{ijt} represent the value for a generic numeric survey item, X, for individual i from community j during year t. The survey weighted mean for survey item X for community j at year t, X_{jt}, is

$$X_{jt} = \left(\sum_{i=1}^{N_{jt}} w_{ijt} X_{ijt} \right) \Bigg/ \left(\sum_{i=1}^{N_{jt}} w_{ijt} \right). \tag{1}$$

Unless otherwise specified, results presented at the community/year level are weighted means.

3 Defining our Scope and Objectives

The parameters of the Data Expo challenge were intentionally vague. The overarching goal was to provide a graphical summary of important features of the data set. Specific questions to help guide exploration were

– What attaches people to their community?
– What are key drivers behind emotional attachment? Are the key drivers all similarly important? What effect does their composition have on attachment?
– How different are the communitites?

Early on we made the decision to favor an exploration of depth rather than breadth. The main reason for this was a practical one. The number of survey items was substantial—nearly 200. Thus, one of the first things we did was read through all survey items and select ones we considered worth further exploration. The survey items we chose are enumerated in table 1.

The years of SoCS, 2008–2010, make up a very interesting and unique time in U.S. history. Some of the major events during these three years included the bursting of the U.S. housing bubble[1], the U.S. auto industry crisis[2] and the Deepwater Horizon oils spill, also known as the BP oil spill[3].

These major events helped direct and define our research objectives and areas of exploration. Each community has its own identity, its own uniqueness, its own story. Thus, we expected each community to respond to these major events in their own way. We anticipated that each community's response would be related to time, geography and demographics. For example, we expected communities' economic responses to the Deepwater Horizon oil spill to be related to their distance from, or their economy's dependence on, the Gulf of Mexico. Other events we expected to have more universal effects. For example, we expected the bursting of the U.S. housing bubble, a significant contributing factor to the 2007–2009 U.S. recession, to have a negative economic effect for all communities, though to varying degrees.

Our exploration objectives were as follows:

– Identify communities that deviate from national trends (i.e., identify outliers)
– Identify drivers of community attachment at both the community and individual levels
– Identify communities with interesting stories
– Identify community responses to significant national events (e.g., the BP oil spill)

4 Temporal Community-Level Effects

To begin our exploration of the SoCS data, we examined city-level index variables to get profiles of each community. The index variables we considered were already included in the data sets; however, we had to aggregate them for each year in each city using the survey weights, as described in equation (1). We were interested in the relative standing of the communities according to these indices rather than their raw indices, so we rescaled the indices to have a minimum of 0 and a maximum of 1 prior to aggregation.

In our analysis we created time plots to investigate community attachment based on indices that we believed would be associated with community attachment. Our initial hypothesis was that economic outlook, aesthetics, passion, loyalty, social capital and civic involvement all impact an individual's attachment to their community; thus, city-level plots should help identify communities warranting further study. In figure 1, we include the most interesting subset of these time plots: the indices for community attachment (top left), economic outlook (bottom left), civic involvement (top right) and aesthetics (bottom right).

Before considering specific communities, we first note discoveries from figure 1. The relative community attachment time plot shows little inter-community variability, as most communities have a relative community attachment index between .6 and .75, with a few exceptions. We also see that community attachment does not appear to exhibit any sort of trend across years. The time plot for relative civic involvement also exhibits little variability, with most relative values for a given year varying by about 0.1. Relative civic involvement does exhibit a trend across years, with a general increase from 2008 to 2009, and then a decrease from 2009 to 2010. Relative civic involvement in 2010 is generally higher than it was in 2008. Relative economic outlook sees the opposite trend of relative civic involvement. That is, almost uniformily, the relative economic outlook index decreases from 2008 to 2009, but then increases from 2009 to 2010. We see the largest inter-community variability with the relative aesthetic index, with

[1] http://en.wikipedia.org/wiki/Timeline_of_the_United_States_housing_bubble#2007

[2] http://en.wikipedia.org/wiki/Effects_of_the_2008%E2%80%9310_automotive_industry_crisis_on_the_United_States

[3] http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill

values ranging from about 0.3 to over 0.8. For most communities, the relative aesthetic index is constant between 2008 to 2010.

From the time plots in figure 1 we also saw interesting patterns emerge at the community level. Specifically, we identified four communities that seemed to have interesting stories: State College, PA, Detroit, MI, Milledgeville, GA and Biloxi, MS. Below we outline the patterns that lead us to select these communities:

- State College, PA had the highest level of community attachment in 2010, and residents felt more attached to State College than they did in 2008.
- Detroit, MI had the second lowest level of community attachment, and the lowest level of perceived economic outlook in 2008 and 2009, with a slight recovery in 2010.
- While the economic outlook for most communities rebounded in 2010, this was not the case for Biloxi, MS and Milledgeville, GA. Both cities saw continued declines in their economic outlook.
- State College, PA has one of the highest indices on the aesthetics index.
- Although the four cities previously discussed differ greatly economically and in overall community attachment, they display similar levels of civic involvement.

In keeping with our "depth over breadth" philosophy, we elected to investigate these four cities further because of their distinctive profiles in figure 1. Throughout the remainder of the paper we will refer to this group of communities as "the four cities."

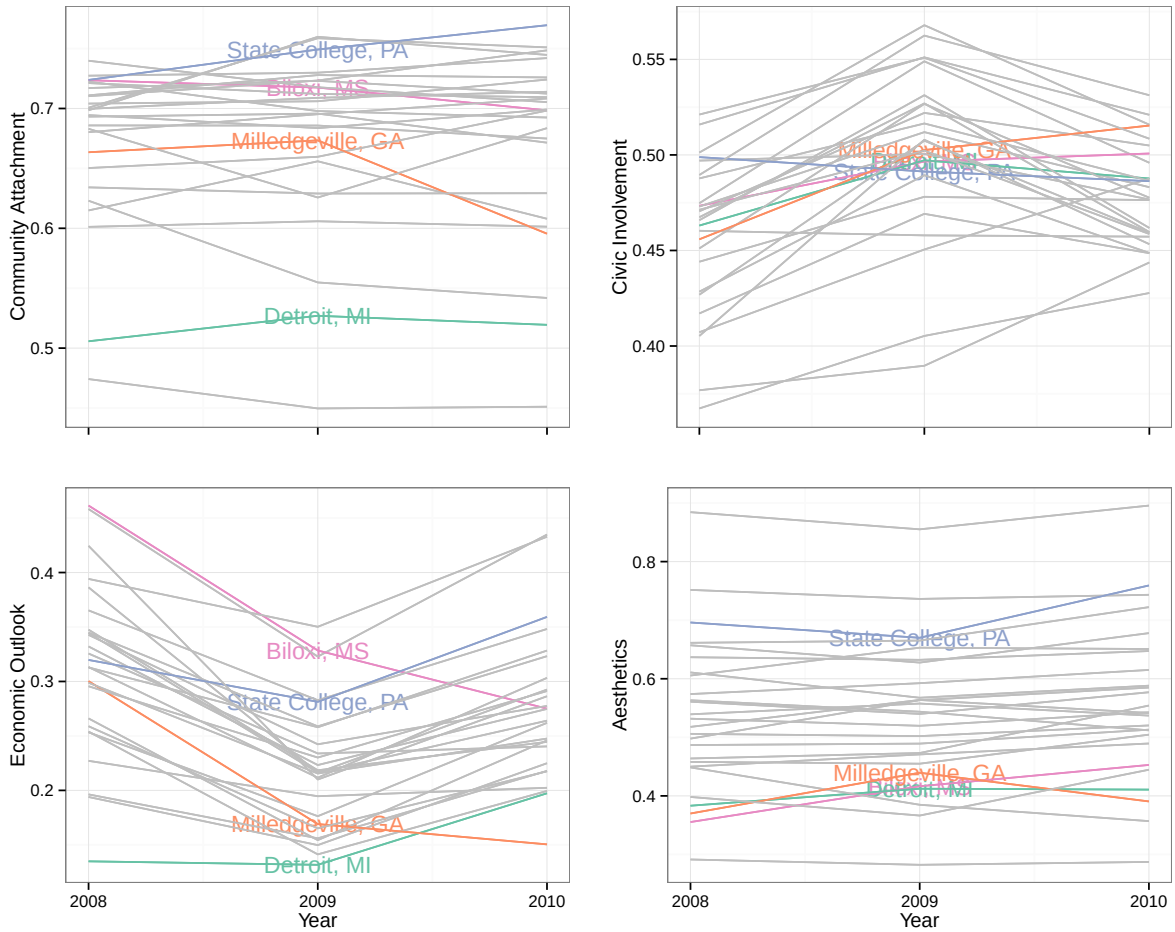

Fig. 1: Time plots of city-level index variables between 2008 and 2010.

5 Economy and Community Attachment

The Knight Foundation wanted to identify the drivers of community attachment. For the four cities of interest we began to explore what index variables were most associated with community attachment. An initial exploration of bivariate relationships found that the economic outlook index held a notable relationship with community attachment in all of the four cities. The historical context of a recovering economy between 2008 and 2010 indicated that we wanted to explore how this relationship evolved over time. This directed us to focus on the relationships between the community attachment and economic outlook indices for individuals within each of the four selected cities in each of the three years.

5.1 Survey-Weighted Binned Scatterplots

To explore the bivariate relationship between the community attachment and economic outlook indices across the years for individuals within each city we needed to develop a graphical approach that captured all elements of the relationship properly. Each of the indices are quantitative so a natural way to visually display the bivariate relationships would be through scatterplots.

In this situation scatterplots are problematic for two reasons: (1) there are a large number of individuals from each city in each year causing issues with overplotting and (2) the observations from all individuals would carry the same visual impact regardless of the survey weight attributed to each individual. Our solution was to use adapted binned scatterplots to display the relationships. A binned scatterplot (Unwin et al. 2006) can be thought of as a two dimensional histogram of sorts, where a binning grid is formed over the x- and y-axes. Then the number of points in each bin is mapped using a continuous color scale to either a shade or saturation, generally with darker shades or more saturation, indicating a higher density of points in a bin.

Binned scatterplots were an effective solution to overplotting, but did not account for the problem with the discrepancy in survey weights for individuals in each bin. So a small modification was made to color the bins using a function to combine the survey weights for the individuals in the bin instead of counting the individuals directly. Simply summing the survey weights within a bin would have been sufficient for coloring bin densities in an individual plot, but since we wanted to investigate many bivariate relationships, in four cities over three seperate years, we needed to scale the summed survey weights so that the color scaling would be consistent across all years. The bin density function used the summed survey weights scaled by the sum of survey weights from all bins. Thus the bin density function for the $(g,h)^{th}$ bin for city j in year t was

$$\beta_{ghjt} = \left(\sum_{i \in B_{ghjt}} w_{ijt} \right) \Bigg/ \left(\sum_{i=1}^{N_{jt}} w_{ijt} \right),$$

where B_{ghjt} is the index set of all individuals in the $(g,h)^{th}$ bin, w_{ijt} is the survey weight of the i^{th} individual in city j in year t, and N_{jt} is the total number of inviduals in city j in year t.

To explore how economic outlook impacts community attachment at the individual level, we examined survey-weighted binned scatterplots for community attachment versus economic outlook indices. We also examined survey-weighted histograms for the community attachment and economic outlook indices, marginally.

5.2 Triangular Distributions

While exploring the relationship between the community attachment (y-axis) and economic outlook (x-axis) indices for the four cities, an interesting triangular pattern (distribution) emerged for all years. The triangular distribution had an economic index dependent lower bound but an economic index independent upper bound. Said another way, if an individual reports a high economic outlook, that individual will

very likely report having high community attachment. If an individual reports a low economic outlook, reporting any level of community attachment is plausible.

The other surprising finding is that the marginal distributions that produce the triangular bivariate distribution are quite varied. To illustrate this, consider figure 2. For 2008, we present survey weighted histograms for the community attachment and economic outlook indices, along with the binned scatterplots described in Section 5.1, for the four cities. State College and Milledgeville have very similar marginal distributions for both the community attachment—skewed left—and economic outlook indices—skewed right. Biloxi has a community attachment index distribution similar to that of State College and Milledgeville—skewed left—but a unimodal and symmetric economic index distribution, indicating the economic profile of Biloxi in 2008 faired better than that of State College and Milledgeville. Detroit's community attachment index distribution differs from the other three communities, as it is not skewed left. It is relatively unimodal and symmetric, indicating the community attachment profile of Detroit is worse than the other three communities. Detroit's economic index distribution is skewed right, similar to that of State College and Milledgeville. The skew of Detroit's economic index distribution is the most severe of the four cities, indicating the worst economic outlook of the four cities.

Despite the different marginal distributions for these four communities, they all exhibit a triangular shaped bivariate distribution. The particulars of the density are specific to each community, but those particulars appear to be constrained by the triangular shape.

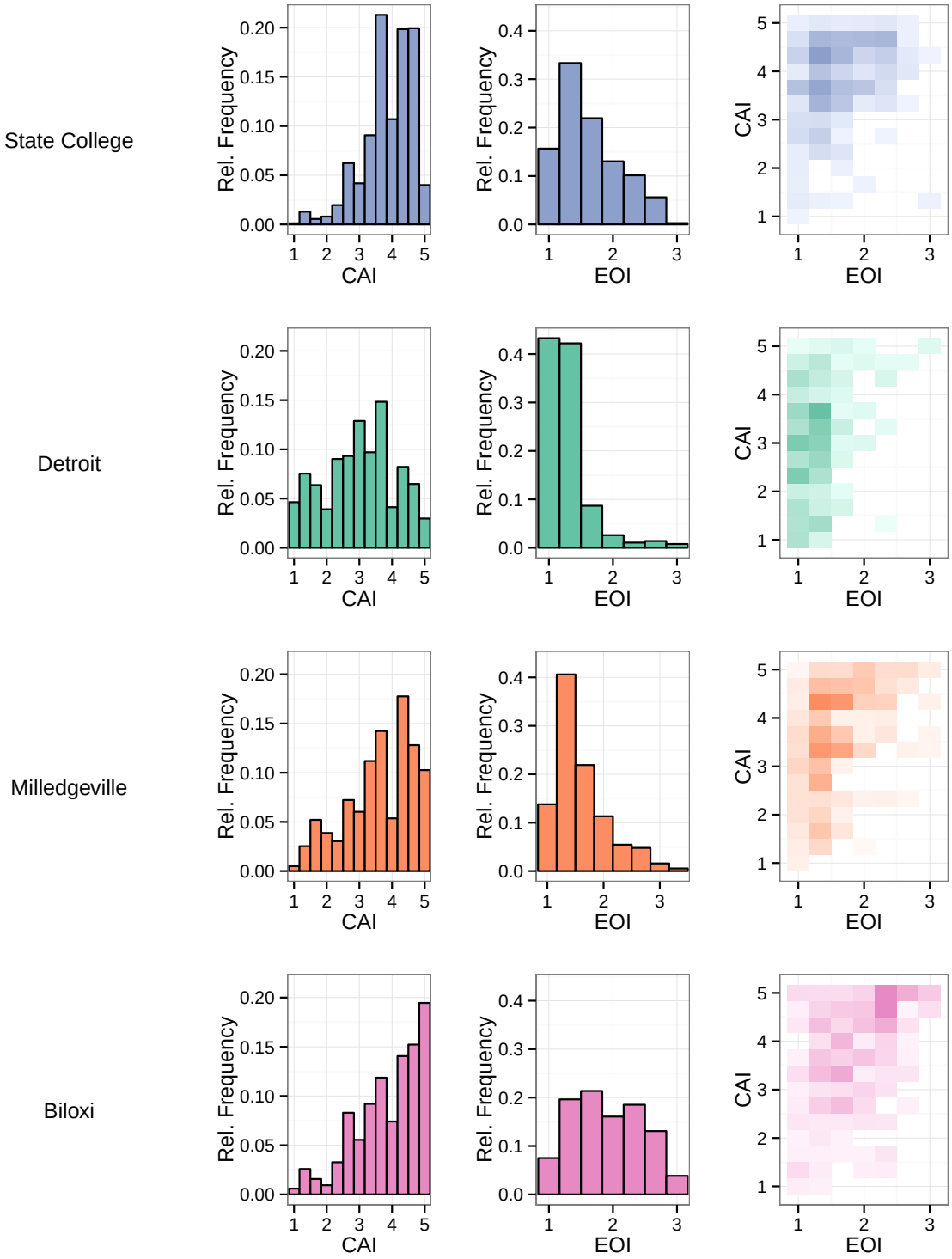

Fig. 2: Histograms of the marginal distributions of the community attachment index (CAI) and economic outlook index (EOI) for the four cities in 2008 along with a binned scatterplot of the bivariate distribution. Interestingly, the bivariate distribution has a triangular shape for each city.

5.3 Tales of Four Cities

5.3.1 State College, PA

In many ways, State College is a model community with respect to several of the measured indices. State College has the highest community attachment in 2010 of all 26 selected communities, along with high scores for the economic outlook, aesthetic and social capital indices.

In figure 3, we focus on the relationship between community attachment and economic outlook between 2008 and 2010. State College was insulated to the country's economic downturn. Though there is a slight down shift in the economic outlook index between 2008 and 2009, it is relatively minor. For reference, the reader is directed to the economic outlook plot of figure 1. Like the rest of the country, we start to see an economic outlook recovery with State College in 2010. The mode of the economic outlook distribution is still low, but the right tail gains density from 2009 to 2010.

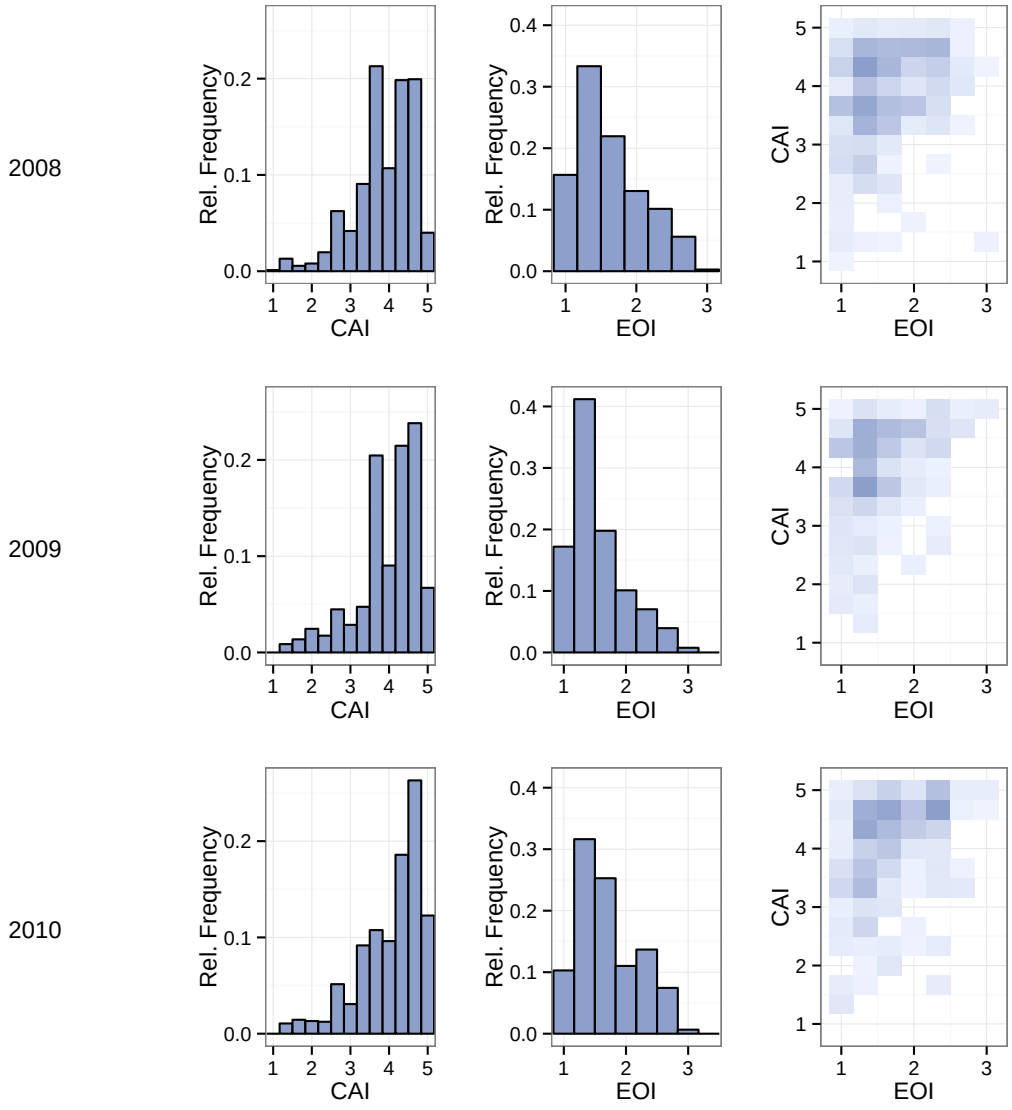

Fig. 3: Histograms of the marginal distribution of the community attachment index (CAI) and economic outlook index (EOI) along with a binned scatterplot of the bivariate distribution for State College, by year.

Between 2008 and 2010, we see that community attachment steadily increases. This was initially surprising to us. State College is the home of Penn State University. Around this time was the arrest of Penn State defensive coordinator Jerry Sandusky, for what is now known as the "Penn State child abuse sex scandal." Our surprise was due to faulty recall, however, as Jerry Sandusky was arrested on November 5th, 2011. Thus any measurable effects due to the Penn State child abuse sex scandal would not be present in these data. It would be interesting to see State College's community attachment information for 2011 and 2012 to see if this event did have any measurable effect.

5.3.2 Detroit, MI

Detroit's economy is heavily rooted in the auto industry, specifically the "Big Three" automakers: Ford, General Motors (GM) and Chrysler. By the latter part of the 2000s, the Big Three were having major financial problems, partially as a result of increasing gas prices. Rising gas prices resulted in increased consumer demand for fuel-efficient cars and diminishing demand for less fuel-efficient SUVs and large pickup trucks. SUVs and large pickup trucks made up a large portion of the Big Three's portfolios. In September of 2008, the financial troubles of the Big Three were national news. It is reasonable to think the adverse regional economic effects were felt in Detroit prior to September of 2008.

To see if this line of thinking is consistent with these data, consider the economic outlook plot of figure 1. All cities see a drop in economic score from 2008 to 2009 except Detroit. We do not believe this is because Detroit was insulated against the economic hardships felt by the rest of the country; but rather, we believe Detroit felt the economic hardships before the rest of the country. This belief is supported by noting Detroit's average economic outlook is the lowest of all 26 considered communities in 2008 and 2009. The narrative that Detroit faced economic hardships partially due to the economic trouble of the Big Three before their troubles became national news in September of 2008 is consistent with these data.

Figure 4 shows a heavily right skewed distribution for economic outlook in 2008 and 2009. Economic outlook improves in 2010, but is still decidedly poor. For all considered years, there is significant spread in the community attachment index. As can be seen by the scatterplots of figure 4, a high economic outlook almost always coincides with high levels of community attachment. Low economic outlook, which is the majority of Detroit, certainly does not imply low community attachment, however.

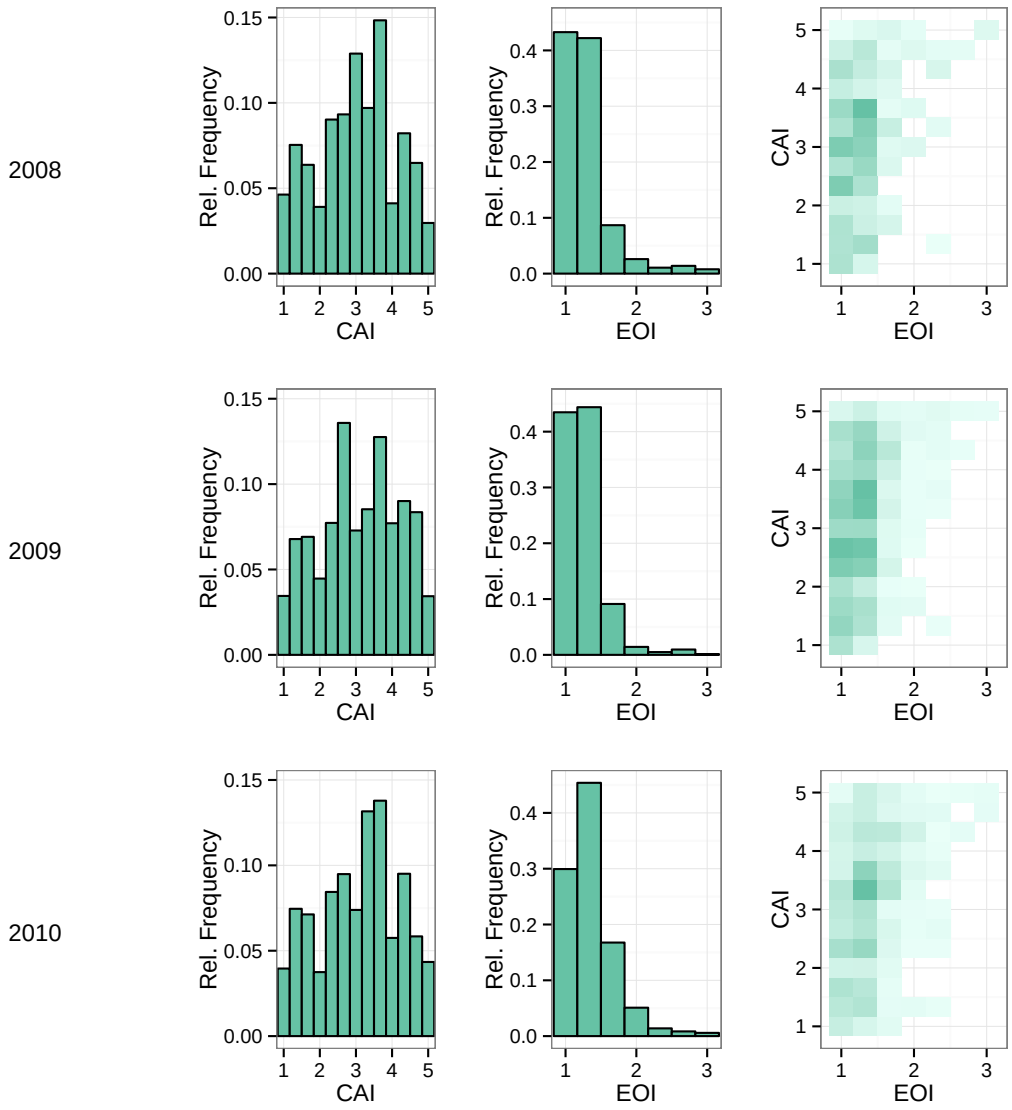

Fig. 4: Histograms of the marginal distribution of the community attachment index (CAI) and economic outlook index (EOI) along with a binned scatterplot of the bivariate distribution for Detroit, by year.

5.3.3 Milledgeville, GA

Milledgeville is a small town in the middle of Georgia that caught our attention in figure 1. We felt Milledgeville deserved further investigation, as it was one of two communities that did not see an economic recovery between 2009 and 2010.

Figure 5 tells the story of a city experiencing increasing levels of economic hardship. As mentioned in Section 5.2, the 2008 community attachment and economic outlook index distributions are quite similar to that of well-off State College. In 2009, we see a diminished economic outlook, similar to that of many cities. Simultaneously, there is a slight increase in community attachment (better seen in figure 1). But when most of the country saw their economic outlook rebound, Milledgeville saw it sink even further. They also saw their community attachment drop by an appreciable amount. The question is why? Why did Milledgeville behave so differently than the rest of the country between 2009 and 2010?

An explanation consistent with these data is that in 2010 the Central State Hospital of Milledgeville, one of the town's largest employers, announced it would close[4]. The decline in economic outlook from 2008 to 2009 can likely be explained by the U.S. financial crisis, felt by the nation. The decline in community attachment and economic outlook from 2009 to 2010 might likely be related to the announced closing of the Central State Hospital.

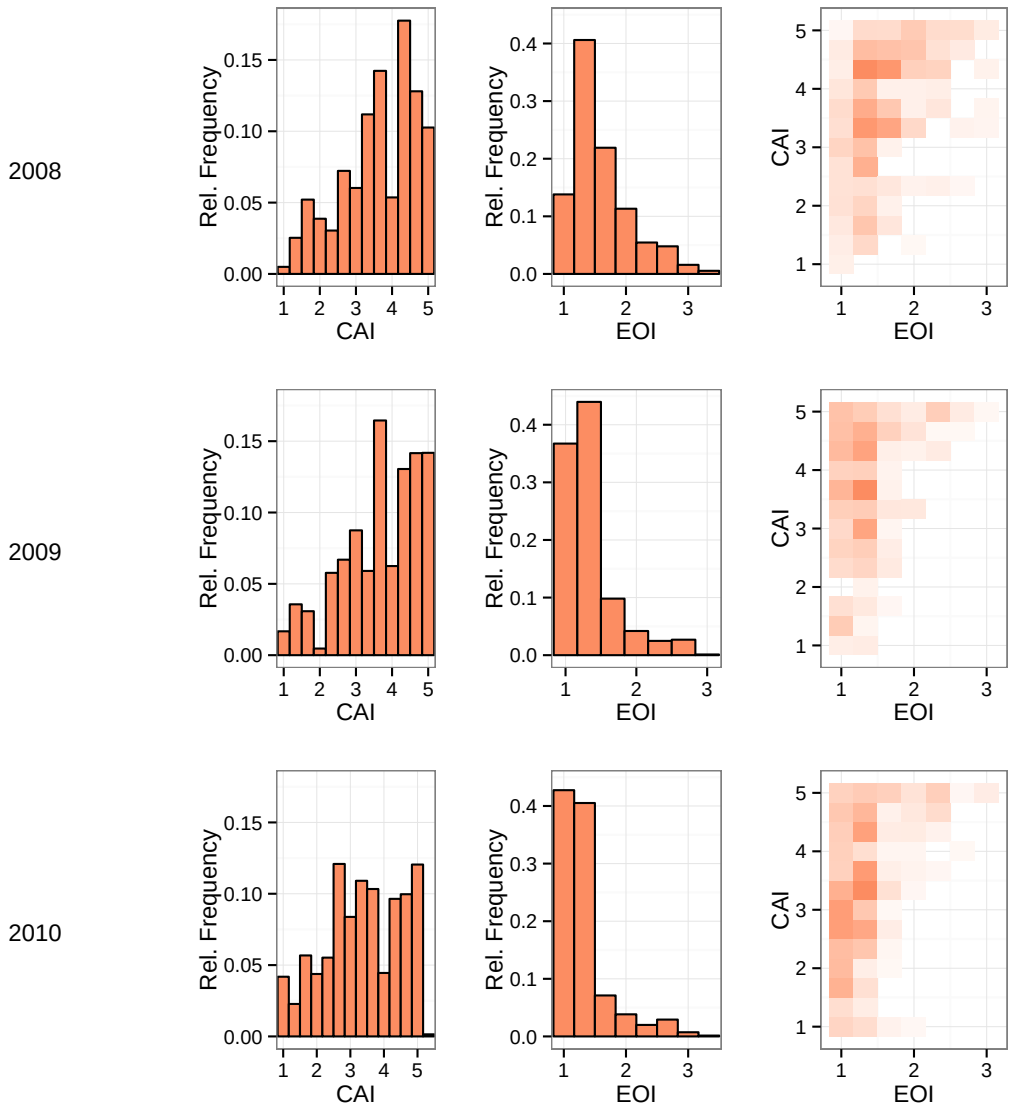

Fig. 5: Histograms of the marginal distribution of the community attachment index (CAI) and economic outlook index (EOI) along with a binned scatterplot of the bivariate distribution for Milledgeville, by year.

5.3.4 Biloxi, MS

Like Milledgeville, the community of Biloxi, MS was the only other community that experienced a decline in economic outlook between 2009 and 2010. Consider figure 6. In 2008, Biloxi had a high community attachment and economic outlook. In fact, Biloxi had the highest economic outlook in 2008 (see figure 1

[4] `http://en.wikipedia.org/wiki/Central_State_Hospital_(Milledgeville,_Georgia)`

for reference). Like most of the 26 cities, Biloxi saw a decline in economic outlook from 2008 to 2009, but a community attachment that remained relatively unchanged. Like Milledgeville but different from the other cities, Biloxi saw a continued decline in economic outlook from 2009 to 2010. Unlike Milledgeville, Biloxi's drop in economic outlook was not accompanied by a substantial drop in community attachment. When we asked why Biloxi did not see an economic rebound like the rest of the nation, the BP oil spill seemed like a plausible explanation.

Biloxi is a community on the Gulf of Mexico whose economy is dependent upon tourism and the seafood industry. The BP oil spill lasted for nearly three months (April 20th through July 15th, 2010). It was imminently clear the BP oild spill would adversely affect communities in the Gulf, especially those dependent upon the industries the Gulf supports. The extent of the BP oil spill damage was, however, a topic of significant speculation. With both of those industries negatively affected in light of the BP oil spill, it seems plausible this event is related to the continued decline in economic scores for Biloxi.

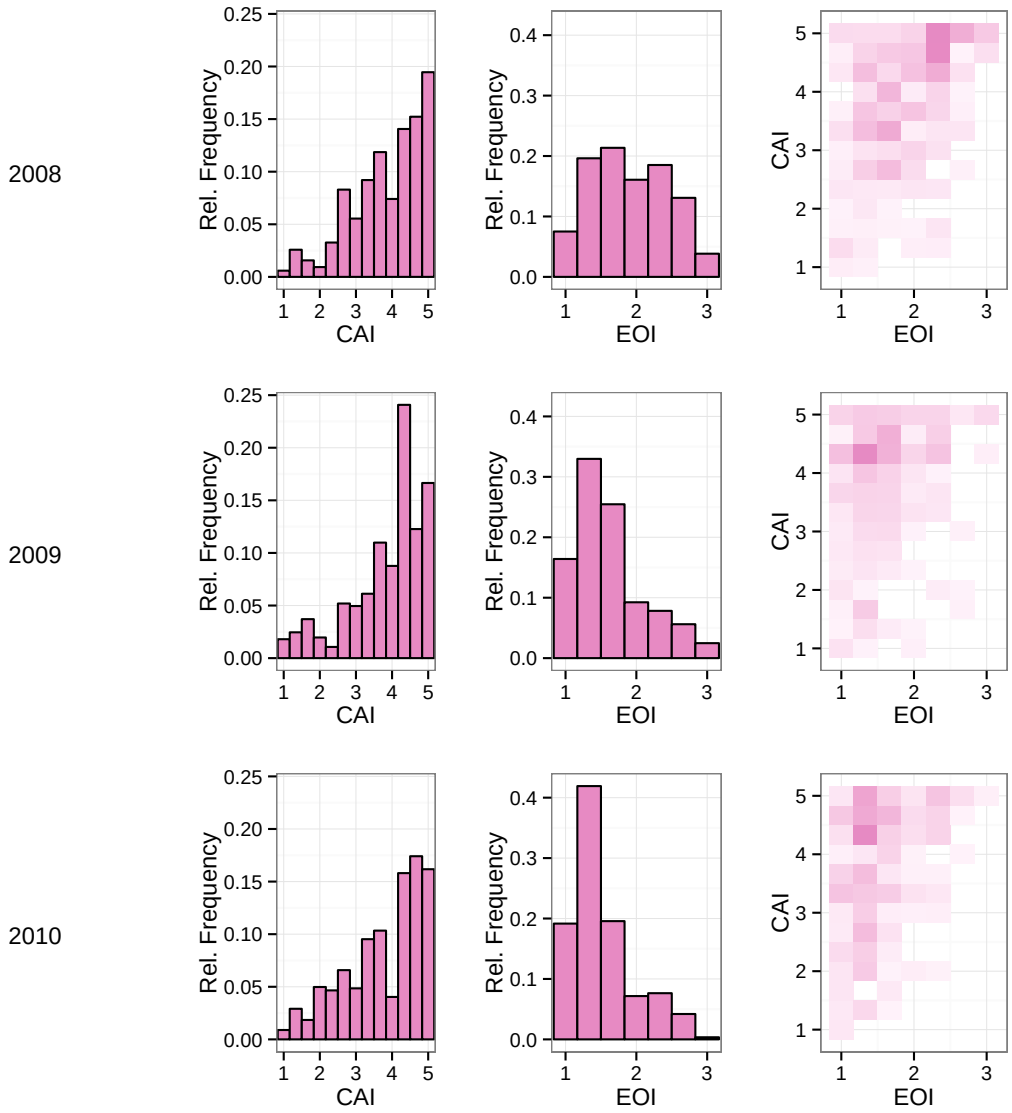

Fig. 6: Histograms of the marginal distribution of the community attachment index (CAI) and economic outlook index (EOI) along with a binned scatterplot of the bivariate distribution for Biloxi, by year.

6 Graphical Sketches by Boz

The following subsections are a loose collection of the interesting findings hidden in these data that did not find a home in our main narrative, the Tale of Four Cities. Luckily, Charles Dickens had a myriad of works. Thus, the following subsections are in the spirit of Charles Dickens' collection of images and short stories, *Sketches by Boz: Illustrative of Every-Day Life and Every-Day People*.

6.1 Economy and Civic Involvement

The old adage is that communities come together in times of hardship. We decided to see how true this was with the 26 communities from the SoCS because these communities faced a period of economic hardship and recovery. To assess this we decided that the survey weighted mean civic involvement index would suffice as a metric of the degree to which a community was actively united. In this context, "Coming together in times of hardship" would imply that as the situation worsened economically we should see a rise in civic involvement.

Figure 7 follows the yearly change in survey weighted mean indices for the economic outlook and civic involvement for all 26 cities. The changes are displayed as a percentage change in index from the previous year so that the "coming together" is measured relative to the community. The four cities from our main narrative are again highlighted in this plot to tie in the familiar economic stories to civic involvement.

Figure 7 is consistent with the adage that communities come together in times of hardship. In the plot we see all communities suffered a worsening economic condition from 2008 to 2009, but almost all also displayed an increase in civic involvement over the same period. Conversely in 2009 to 2010 all communities with the exception of Biloxi and Milledgeville experienced an improving economic outlook and nearly all of them saw a decrease in civic involvement. Perhaps the old adage should be updated to "communities come together in times of hardship, but when times are good you are on your own!"

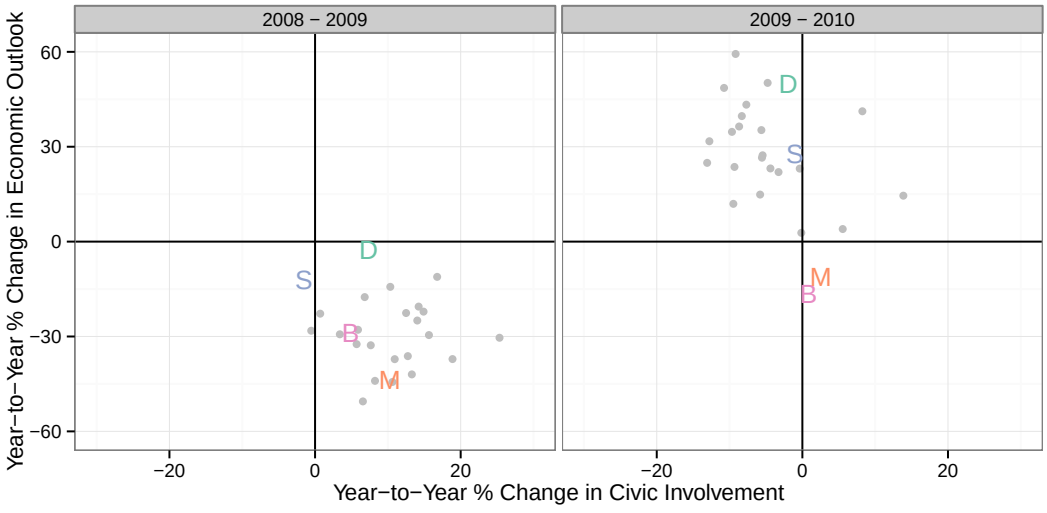

Fig. 7: Scatterplots of year-to-year percent change in the economic outlook index against year-to-year percent change in the civic involvement index for 2008 to 2009 (left) and 2009 to 2010 (right). All 26 communities are included in the plot, but the four cities are plotted using the first letter of their name (S = State College, D = Detroit, B = Biloxi, and M = Milledgeville).

6.2 Friends or Family?

We investigated how the proportion of one's friends and family living in their community is related to community attachment. We leveraged and paired two questions with individual community attachment scores. Those questions were:

- How much of your family lives in the area?
- How many of your close friends live in your community?

For each question, the possible answers were: "A Few", "Some", "About half", "Most" and "All or nearly all." We partitioned all SoCS participants into the 25 groups formed by the factorial design of the five question levels. For each group, we calculated the survey-weighted average community attachment score. A heat plot of average community attachment scores for each group is displayed in figure 8.

A relatively clear story is told by figure 8. Namely, the more close friends that live in your area, the higher your community attachment will likely be. The same does not hold for family. This is best illustrated by focusing on the edges of figure 8. For all levels of family in the area, we see high levels of average community attachment for individuals with "All or nearly all" of their close friends in the area. For all levels of family in the area, we see low levels of average community attachment for individuals with only "A few" close friends in the area. For any level of family in the area, we generally see an increase in community attachment as the proportion of close friends in the area increases. Though not necessarily a causal relationship, this plot suggests a possible way to increase the community attachment of individuals in a community is by supporting activities and events that enable people to meet each other and make friends.

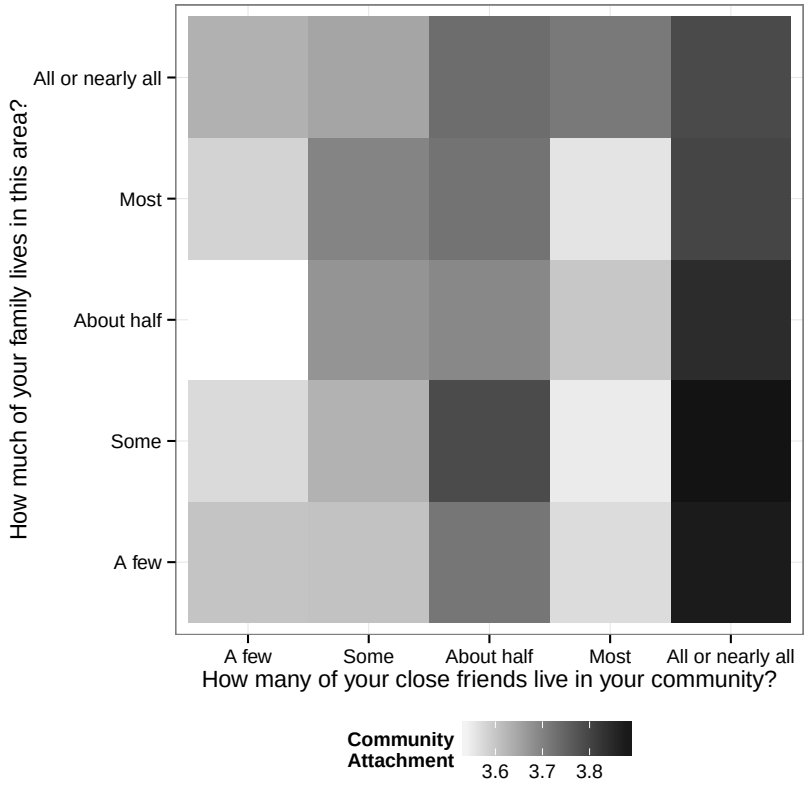

Fig. 8: Heat plot of average community attachment scores for the combination of answers to the questions: (1) How much of your family lives in the area? and (2) How many of your close friends live in your community?

6.3 Acceptance and Community Attachment

If a community is more accepting of racial and ethnic minorities, will it likely have higher community attachment than a community that is less accepting? What about their acceptance of immigrants? Or how about gay and lesbian people? We sought evidence related to these three questions.

Survey weighted averages over all three years for each community were calculated for the community attachment score along with answers to the following three questions:

- How accepting are you of gay and lesbian people?
- How accepting are you of immigrants?
- How accepting are you of racial and ethnic minorities?

Figure 9 presents scatterplots of community attachment score versus these three questions. A couple things are noteworthy. The first is that the triangular distribution described in section 5.2 is present in all three plots, though most markedly in the leftmost plot of figure 9. Communities that express high levels of acceptance for gay and lesbian people, immigrants, or racial and ethnic minorities tend to have high levels of community attachment. For communities that express low levels of acceptance for gay and lesbian people, immigrants, or racial and ethnic minorities, all levels of community attachment are reasonable within the range of these data. The second interesting point is that the variability of acceptance levels is smallest with acceptance of racial and ethnic minorities, larger with acceptance of immigrants and largest with acceptance of gay and lesbian people. This suggests acceptance of gay and lesbians was more heterogeneous than acceptance of immigrants or racial and ethnic minorities, between 2008 and 2010.

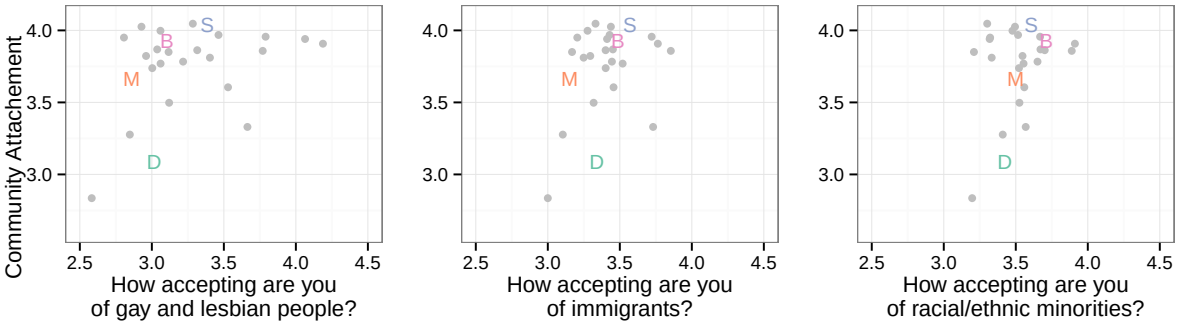

Fig. 9: Scatterplot of the community attachment index against the answers to the acceptance questions. All 26 communities are included in the plot, but the four cities are plotted using the first letter of their name (S = State College, D = Detroit, B = Biloxi, and M = Milledgeville).

7 Summary of Findings

Many insights were gained throughout the exploration of the SoCS data. An enumeration of our findings are:

- Community-level, survey weighted community attachment averages are surprisingly constant over the years.
- Community-level, survey weighted economic outlook averages exhibit an almost universal drop from 2008 to 2009, and an almost universal rebound from 2009 to 2010.
- A triangular distribution between community attachment and economic outlook is present for virtually all cities for all years. Given an individual has a high economic outlook, they very likely will also have a high community attachment. Given an individual has a low economic outlook, all levels of community attachment are possible.

– Milledgeville, GA and Biloxi, MS saw continued declines in their respective economic outlook indices, deviating from the national trend. Upon further investigation, we were able to find plausible explanations for these deviations. One of the largest employers in Milledgeville, Central State Hospital, announced its closing in 2010. The BP oil spill adversely affected critical economic industries for Biloxi, plausibly contributing to the decreasing economic outlook.
– Civic involvement and economic outlook are highly negatively related. Economic outlook almost universely declines from 2008 to 2009, while civic involvement almost universely increases. When economic outlook rebounds in 2010, civic involvement almost universely decreases.
– Friendship appears to be a better predictor of community attachment than family. Having many friends nearby is linked with high levels of community attachment. Having many family members nearby is not.
– Given a community has a high level of acceptance for gay and lesbian people, immigrants, or racial and ethnic minorities, that community likely has a high level of community attachment. Given a community has a low level of acceptance, all levels of community attachment are likely.

8 Tools

All of the computation necessary to create the graphics found on the Data Expo poster, and in this paper, was performed in R (R Core Team 2013), and all graphics were created using `ggplot2` (Wickham 2009). We also used `plyr` (Wickham 2011) to help perform data aggregation and manipulation, and we used `gridExtra` (Auguie 2012) to create custom layouts of our `ggplot` graphics. Other packages were also used, but these were by far the most helpful (we refer the reader to our R script for the full list). We used Adobe InDesign to create the poster, and typeset this paper using LaTeXand the R package `knitr` (Xie 2013).

References

Auguie, B. (2012). gridExtra: functions in Grid graphics. R package version 0.9.1.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Unwin, A., Theus, M., and Hofmann, H. (2006). Graphics of large datasets: visualizing a million. Springer.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer New York.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. Journal of Statistical Software, 40(1):1–29.

Xie, Y. (2013). Dynamic Documents with R and knitr. Chapman and Hall/CRC.

Table 1: Selected survey items for further exploration.

	Variable Names	SoCS 2008	SoCS 2009	SoCS 2010	Note
1	id	CASE	CASE	CASEID	*city id
2	city	QSB	QSB	QSB	Specific Knight Communities can be identified by variable QSB.
3	qs3	QS3	QS3	QS3	*this number is assigned to a city (for instance 106 and 107 belong to Aberdeen, SD)
4	zip	QS3A	QS3A	QS3A	*zip code
5	cityType	QS4	QS4	QS4	*city category
6	citySubTypeWord	QS5	QS5	QS5	*in 08, QS5 is the word version of QS5_2, which we will use to convert QS5 in 2009 and 2010
7	citySubTypeNum	QS5_2	QS5_2	QS5_2	*city sub category rep by number
8	comSat	QCE1	QCE1	QCE1	*community satisfaction
9	proud	Q3A	Q3A	Q3A	*I am proud to say I live in the place I live
10	comp5YearPast	Q6	Q6	Q6	*compare your place to live now to it 5 years ago
11	comp5YearFuture	Q6A	Q6A	Q6A	*how do you think your place of residence in 5 years will compare to it now?
12	affordHousing	Q7D	Q7D	Q7D	*availability to affordable housing
13	jobAvail	Q7E	Q7E	Q7E	*availability to jobs
14	imAccept	Q8B	Q8B	Q8B	*how accepting are you of immigrants
15	raceAccept	Q8C	Q8C	Q8C	*how accepting are you of racial and ethnic minorities
16	gayAccept	Q8E	Q8E	Q8E	*how accepting are you of gay and lesbian people
17	econCondNow	Q9	Q9	Q9	*what do you think is the economic status now
18	econCondFuture	Q10	Q10	Q10	*what do you think will be the economic status in the future
19	employStatus	Q11	Q11	Q11	*what is your employment status
20	incomeSat	Q15	Q15	Q15	*does your job provide you with enough income to support your family
21	crimeNow	Q19	Q19	Q19	*crime level in your community today
22	crimePast	Q20	Q20	Q20	*how has crime level changed in your community in the last year
23	volunteerWork	Q22A	Q22A	Q22A	*Performed local volunteer work for any organization or group
24	comWork	Q22D	Q22D	Q22D	*Worked with other residents to make change in the local community
25	closeFriends	Q24	Q24	Q24	*How many of your close friends live in your community?
26	closeFam	Q25	Q25	Q25	*How much of your family lives in this area?
27	age	QD1	QD1	QD1	*how old are you
28	duration	QD2	QD2	QD2	*how long have you lived in this community
29	permanent	QD2A	QD2A	QD2A	*do you live here permanently
30	maritalStatus	QD6	QD6	QD6	*marital status
31	eduMax	QD7	QD7	QD7	*education
32	ownRent	QD8	QD8	QD8	*own or rent
33	income	QD9	QD9	QD9	*income
34	hispanic	QD10	QD10	QD10	*are you of hispanic origin
35	race	QD111	QD111	QD111	*Which of these groups best describes your racial background?
36	svywt	WEIGHT	WEIGHT	WEIGHT	*survey weight
37	projwt	PROJWT	PROJWT	PROJWT	*projection weight
38	passion	PASSION	PASSION	PASSION	*passion
39	loyalty	LOYALTY	LOYALTY	LOYALTY	*loyalty
40	basicServ	BASIC_SE	BASIC_SE	BASIC_SE	*basic services
41	leadership	LEADERSH	LEADERSH	LEADERSH	*leadership
42	education	EDUCATIO	EDUCATIO	EDUCATIO	*education
43	safety	SAFETY	SAFETY	SAFETY	*safety
44	aesthetic	AESTHETI	AESTHETI	AESTHETI	*aesthetics
45	economy	ECONOMY	ECONOMY	ECONOMY	*economy
46	socialOff	SOCIAL_O	SOCIAL_O	SOCIAL_O	*social offerings
47	civicInv	INVOLVEM	INVOLVEM	INVOLVEM	*civic involvement
48	openness	OPENNESS	OPENNESS	OPENNESS	*openness
49	socialCap	SOCIAL_C	SOCIAL_C	SOCIAL_C	*social capital
50	domains	DOMAINS	DOMAINS	DOMAINS	*domains
51	comOff	COMMUNIT	COMMUNIT	COMMUNIT	*community offerings
52	comAttach	CCE	CA	CCA	*community attachment
53	comAttachGrpSmall	CCEGRP	CAGRP	CCAGRP	*community attachment group (few groups)
54	comAttachGrpBig	CCE_ENGA	CA_ATTAC	CCA_ATTA	*community attachment group (many groups)
55	gender	INTRO1	INTRO1	GENDER	*gender (we need to substring stuff for intro1 (08 and 09))
56	year	YEAR	YEAR	YEAR	*the year