# Introduction

This document describes the features of the compiled and cleaned version of the Crowd Counting Consortium (CCC) data on U.S. protest events from the project’s first phase, which ran from January 2017 through December 2020.

The project started by counting the Women’s March of January 21, 2017.[[1]](#footnote-0) It expanded immediately thereafter, seeking information on all protests, rallies, and marches in the United States. For February and March 2017, CCC largely relied on Google searches using terms like “protest” or “rally.” Starting in April 2017, however, Count Love — another protest event analysis project — shared with CCC the results of its daily web crawl of more than 3,400 local and national news outlets. This partnership ended in January 2021 when Count Love stopped running its crawler and CCC set up its own version using Count Love's [open-source example](https://github.com/count-love/crawler). Most CCC protest information came from traditional news sources, with non-systematic use of social media — especially Twitter and, to a lesser degree, Facebook. CCC also had a mechanism for the general public to submit protest information, and that pathway contributed a small share of the data. At times of major, multi-location protests like the Women’s March, CCC also made use of organizational websites. The CCC data included reports of upcoming protests as well as ex post coverage of events. The project relied upon both volunteer and paid research assistants.

In this compiled and augmented version of the data, columns that appear in some of the CCC Google Sheets from that period but are specific to certain months or macro-events have been dropped; some columns have been added or modified to facilitate mapping and analysis (e.g., FIPS codes); and events occurring outside the United States and U.S. territories have been omitted.

# Variable Descriptions

## date

Date of event in YYYY-MM-DD format. When an event spans multiple days, the start date is used.

## locality

Name of the locality in which the event took place. Labeled *CityTown* in the Google Sheets.

## state

Two-letter U.S. postal abbreviation for the state or U.S. territory in which the event took place. Labeled *StateTerritory* in the Google Sheets, but the version here has attempted to correct errors in the original.

## location\_detail

Where available, text giving additional details on the location(s) within the city or town where the action took place. Labeled *Location* in the source data.

## online

Binary indicator for online-only events. 1 = online, 0 = in-person. Generated from location and event type information in the source data.

## type

Type(s) of protest action (e.g. march, protest, demonstration, strike, counter-protest, sit-in), separated with semicolons or commas when more than one. *EventType* in source data.

## macroevent

A unique id that associates a counter-protest with the protest event it countered. These strings are composed of a date, a location, and something about the nature of the event, all separated by hyphens (e.g., "20220624-phoenix-abortion"). In most cases, these ids will uniquely identify pairs of events. In cases where the counter-protest is itself countered, however—e.g., a community defense action in response to a protest targeting an LGBTQ+ pride festival—the additional events are given the same id, so these clusters will sometimes include three or more events.

NOTE: This field was created in 2021, and it was only retroactively generated for events associated with the Black Lives Matter protest wave that began in May 2020. So, for CCC Phase 1 data, it is only useful for analysis of that specific subset.

## actors

The organization(s) that organized the protest event (e.g. Women's March, Greenpeace, etc.), and/or the type of people participating (e.g., students, nurses). Usually separated by semicolons, sometimes by commas. Labeled *Actor* in the Google Sheets.

## claims

A phrase or phrases describing what the event was about, based on the claims or demands the participants made (e.g. for women's rights, anti-Muslim Ban, against racism, etc.), as summarized by coders. Labeled *Claim* in the Google Sheets. Phrases are separated by semicolons in some records, commas in others.

## valence

Political valence of the event. Labeled *Pro(2)/Anti(1)* in the source data.

2 = pro-Trump

1 = anti-Trump

0 = other/neither

## issues

String of semicolon-separated tags identifying political issues (or themes) associated with the event (e.g., "democracy; women's rights" for events associated with the 2017 Women's March). These are generated after data compilation by running a series of regular expressions over the claims description text.

## size\_text

Words or phrases that journalists or eyewitnesses used to describe the size of the crowd at the protest event (e.g., “more than 100”, “dozens”, “about 50”)—or, in cases where no text descriptions were found but other sources gave information about crowd size, the alternative source used: “count pic” for photograph(s), “count vid” for video(s), “count FB” for Facebook, or “eyewitness” for eyewitness. *EstimateText* in the source data.

## size\_low

Lowest crowd size reported in, or estimated from, *size\_text*. *EstimateLow* in source data.

To convert vague text counts to numbers, the following three rules are always used for both *size\_low* and *size\_high* (see below).

1. For phrases with fudge words (e.g., “about”, “nealy” “approximately”, “maybe”), ignore the fudge word and treat the number that follows it as the count.
2. When unspecific multiples are given, assume the multiple is 2 (e.g., “hundreds” becomes 200, “thousands” becomes 2,000).
3. Assume “several” means 3 (e.g., “several dozen” becomes 36, “several thousand” becomes 3,000).

## size\_high

Highest crowd size reported in, or estimated from, *size\_text*. *EstimateHigh* in source data.

When vague text counts are converted to numbers, the high and low estimates are assumed to be equivalent. For example, if the only information available about crowd size is a single report describing it as “hundreds”, 200 would be assigned to both *size\_low* and *size\_high*. If, however, one source described the crowd size as “hundreds” and another as “about 500”, then *size\_low* would still be 200, but *size\_high* would be 500.

## size\_mean

The mathematical average of *size\_low* and *size\_high*, rounded up to the nearest integer.

## size\_cat

Ordered categorical indicator of crowd size, representing orders of magnitude and derived from *size\_mean*.

0 = unknown

1 = 1-99 (tens)

2 = 100-999 (hundreds)

3 = 1,000-9,999 (thousands)

4 = 10,000+ (tens of thousands)

## arrests

Text, sometimes specifying the count of protesters reportedly arrested (e.g., “5”), sometimes a phrase indicating ambiguity about that count (e.g., "more than 5", "unclear", “unspecified”). *ReportedArrests* in the source data.

## arrests\_any

Binary indicator for whether or not any arrests occurred, derived from *arrests*. 1 = yes, 0 = no.

## injuries\_crowd

Text, sometimes giving a count of protesters reportedly injured (e.g., “5”), sometimes a phrase indicating ambiguity about that count (e.g., "more than 5", "unclear", “unspecified”). Labeled *ReportedParticipantInjuries* in the source data.

## injuries\_crowd\_any

Binary indicator for whether or not any protesters were reportedly injured, derived from *injuries\_crowd*. 1 = yes, 0 = no.

## injuries\_police

Text, sometimes giving a count of police officers reportedly injured (e.g., “5”), sometimes a phrase indicating ambiguity about that count (e.g., "more than 5", "unclear", “unspecified”). Labeled *ReportedPoliceInjuries* in the source data.

## injuries\_police\_any

Binary indicator for whether or not any police officers were reportedly injured, derived from *injuries\_police*. 1 = yes, 0 = no.

## property\_damage

Text, usually a binary indicator for whether or not any property damage occurred, sometimes a count or other number of unclear meaning, sometimes something else. Labeled *ReportedPropertyDamage* in the source data.

## property\_damage\_any

Binary indicator for whether or not protesters reportedly caused any property damage, derived from *property\_damage*. 1 = yes, 0 = no.

## chemical\_agents

Binary indicator for whether or not police or other state security forces used tear gas or other chemical irritants, such as pepper spray or pepper balls, on protesters. Labeled *TearGas* in the source data, but only available for May–December 2020.

## source\_n

URL of nth source, or description where the source is not a web page (e.g., “eyewitness”, “correspondence”).

## notes

Miscellaneous additional information about the event as noted by the coder. *Misc* or *Misc.* in the source Google Sheets.

## lat

Latitude of locality in which the event took place, as resolved by Google Maps Geocoding API. Note that this is not based on address or landmark-level information where that is given, only on the name of the city or town.

## lon

Longitude of locality in which the event took place, as resolved by Google Maps Geocoding API. Note that this is not based on address or landmark-level information where that is given, only on the name of the city or town.

## resolved\_locality

Name of the locality in which the event occurred, as resolved by running the city or town name and state abbreviation through the Google Maps Geocoding API.

## resolved\_state

Postal abbreviation of the state or territory in which the event occurred, as resolved by the Google Maps Geocoding API.

## resolved\_county

Name of the county in which the event occurred, as resolved by running *resolved\_locality* and *resolved\_state* through the Google Maps Geocoding API..

## fips\_code

Five-digit FIPS code for the county (or LA parish or AK borough or independent city or DC or U.S. territory) given in *resolved\_county*. See 'data-compilation/fips\_for\_county\_function.r' for details on how these are generated using the 'tigris' package and some custom code to handle various exceptions.

NOTE: When you load the data from the stored .csv, you will probably need to add leading zeros back to FIPS codes that have them, because your software will probably read that column as integers instead of strings. In R, you could do this with `ifelse(nchar(fips\_code) == 4, paste0("0", fips\_code), fips\_code)`.

1. Erica Chenoweth and Jeremy Pressman, “This is what we learned by counting the women’s marches,” *Washington Post*, February 7, 2017, <https://www.washingtonpost.com/news/monkey-cage/wp/2017/02/07/this-is-what-we-learned-by-counting-the-womens-marches/>; Jeremy Pressman, Erica Chenoweth, Tommy Leung, L. Nathan Perkins, and Jay Ulfelder, “Protests Under Trump, 2017-2021,” *Mobilization* 27:1 (2022), 13-26; and Kaveh Waddell “The Exhausting Work of Tallying America’s Largest Protest.” *The Atlantic*, January 23, 2017. [↑](#footnote-ref-0)