PERSPECTIVE NATIONAL

For The U.S. Census, Keeping Your Data Anonymous And Useful Is A Tricky Balance

Updated August 2, 2021 · 12:17 PM ET
Heard on Weekend Edition Saturday


HANSI LO WANG

4-Minute Listen

PLAYLIST Download Transcript

People holding umbrellas walk through New York City's Times Square in 2019. The U.S. Census Bureau plans to change how it protects the confidentiality of people's information in the detailed demographic data it produces through the 2020 count.

Mary Altaffer/AP

As the country waits for more results from last year's national head count, the U.S. Census Bureau is facing an increasingly tricky balancing act.

How will the largest public data source in the United States continue to protect people's privacy while also sharing the detailed demographic information used for redrawing voting districts, guiding federal funding, and informing policymaking and research for the next decade?

Concerns have been brewing among census watchers about how the bureau will strike that balance, beginning with the redistricting data it's on track to put out by Aug. 16.

That release is expected to be the first set of 2020 census statistics to come with controversial new safeguards that bureau officials say are needed to keep people anonymous in publicly available data and prevent the exploitation of their personal information. But based on early tests, many data users are alarmed that the new privacy protections could render some of the new census statistics useless.

Ahead of the data's release, the bureau has warned users about how the privacy protections will make some neighborhoods look "fuzzy." The new data may show some blocks with "unusually large" households, children appearing to live alone or "occupied" housing units in areas where the population count is zero, according to a recent blog post by the bureau's acting director, Ron Jarmin.

"Though unusual, situations like these in the data help confirm that confidentiality is being protected," Jarmin said, adding that the "fuzziness disappears" as the individual blocks are grouped together.

NATIONAL

Here's How The 1st 2020 Census Results Changed Electoral College, House Seats

The state of Alabama filed a federal lawsuit to try to block the bureau from putting these new protections in place. In July, a three-judge court put the case on hold after it rejected Alabama's request for an emergency court order, allowing the bureau to continue with its new privacy protection plans.

Still, the topic may resurface in the courts after the bureau releases the data many state and local redistricting officials need to prepare for upcoming elections.

Here's what else you need to know:

Why does the Census Bureau have to protect people's privacy?

Under current law, the federal government is not allowed to release personally identifiable information from the census until 72 years after it's gathered for the constitutionally mandated tally. The bureau has relied on that promise of confidentiality to get many of the country's residents to volunteer their information once a decade, especially among people of color, immigrants and other historically undercounted groups who may be unsure about how their responses could be used against them.

But it is becoming harder for the bureau to uphold that pledge and continue releasing statistics from the census. Advances in computing and access to voter registration lists and commercial data sets that can be cross-referenced have made it easier to trace purportedly anonymized information back to an individual person.

NATIONAL

Immigration Hard-Liner Files Reveal 40-Year Bid Behind Trump's Census Obsession

For a way out of this conundrum, the bureau has been building a new privacy protection system based on a mathematical concept known as differential privacy. Invented at Microsoft's research arm, it has served as a framework for privacy measures in smaller Census Bureau projects, as well as at some tech companies.

"Differential privacy is in every iPhone and every iPad," says Cynthia Dwork, a computer scientist at Microsoft Research and Harvard University who co-invented differential privacy. "That may have a larger scale than the number of respondents to the U.S. decennial census, but there's a totality and commitment to privacy that's different here" with the bureau's plans for 2020 census data, Dwork adds.

How has the bureau protected people's privacy in past census data?

For decades, the bureau has stripped away names and addresses from census records before turning them into anonymized data. That information is broken down by race, ethnicity, age and sex to levels as detailed as a neighborhood.

NATIONAL

COMIC: How Your State Wins Or Loses Political Power Through The Census

But even in a sea of statistics, certain households — particularly those in the minority of a community — can stick out because they live in isolated areas or have other distinctive characteristics that could make it easier to reveal who they are.

As part of additional privacy protections over the years, the agency has withheld some data tables, and sometimes particular cells within tables, from the public in the past. The bureau has also added "noise" — or data for fuzzing the census results — to certain tables before releasing them. Beginning with data from the 1990 count, it has used a technique called "swapping" to switch out data about certain households with those from different neighborhoods.

What prompted the bureau to choose differential privacy to protect 2020 census data?

In 2016, researchers at the bureau began conducting internal experiments to test the strength of the privacy protections used for 2010 census data, and based on the results, agency officials concluded they can no longer rely on data swapping.

NATIONAL

How 26 People In The Census Count Helped Minnesota Beat New York For A House Seat

Using a fraction of the census data the bureau released a decade ago, the researchers were able to reconstruct a complete set of records for every person included in the

2010 census numbers. Then, after cross-referencing that reconstructed data with records bought from commercial databases, they were able to re-identify 52 million people by name, according to a court filing by John Abowd, the bureau's chief scientist. In a worst-case scenario, the bureau's researchers estimated, attackers with access to more commercial data could unmask the identities of as many 179 million people, or 58% of the population included in the 2010 census.

To try to better protect people's privacy for the 2020 census, the bureau announced in 2017 plans to create a new system, based on differential privacy, that officials say allows them to add the least amount of noise needed to preserve privacy in most of the released data and balance confidentiality and usability.

"Obviously, you know, it's not the easiest thing to do," Jarmin, the bureau's acting director, said at the Population Association of America's annual meeting in May, adding that the bureau decided against data swapping and withholding certain tables as alternative safeguards. "To achieve a similar level of privacy protection with those sort of traditional methods, I think, would have produced a product that was even ... less useful for data users than what we're contemplating right now."

How will differential privacy affect 2020 census data?

The bureau says no noise was added to protect people's privacy in the new state population numbers, including those used to reallocate congressional seats and Electoral College votes, as well as numbers for Washington, D.C., and Puerto Rico. The bureau is also planning to release the total number of housing units in each census block, as well as the number of prisons, college dorms and other group-living quarters in each block, without privacy protections.

NATIONAL

Stuck At 435 Representatives? Why The U.S. House Hasn't Grown With Census Counts

But it remains unclear exactly how the bureau's differential privacy plans will affect other new redistricting data that is expected out by Aug. 16, including population numbers and demographic details about counties, cities and other smaller areas.

It depends on the amount of noise the bureau chooses to add and how it tries to smooth out the effects of that through an operation the bureau calls "post-processing." In June, bureau officials announced their final privacy settings for the new redistricting data, which the bureau said will "lead to lower noise infusion" than the amount of noise that was added to test files the bureau released in April to demonstrate the potential effects of its privacy plans. The revised algorithm "ensures the accuracy of data necessary for redistricting and Voting Rights Act enforcement," the bureau said in its announcement.

During a webinar in July, Michael Hawes, the bureau's senior adviser for data access and privacy, noted that compared to the privacy settings applied to the April test files, the final settings resulted in "a step backwards" in terms of the "accuracy" of data about census blocks, which are similar to city or neighborhood blocks and make up the smallest level of geography for which the bureau provides demographic information.

"This will help protect locational privacy," Hawes said. "And because we took a little bit of accuracy away from blocks, that allowed us to put a little bit more accuracy on [census] block groups, [census] tracts, counties and other sort of geographies that we believe are more useful for data users."

At the census block level, users of 2020 census redistricting data may see "unusually large" households, children appearing to live alone or "vacant" housing units in an area where the population count is "greater than zero," Jarmin, the bureau's acting director, warned in a July blog post.

"Instead of looking for precision in an individual block, we strongly encourage data users to aggregate, or group, blocks together," Jarmin said in the post. "As blocks are grouped together, the fuzziness disappears. And when you step back with more blocks in view, the details add together and make a sharp picture."

In July, the bureau announced that at the same time it releases new redistricting data, it plans to put out the new test files users need to analyze the data's usability. The test files will allow data users to compare the effects of the bureau's finalized plans with those of simulated settings for earlier test files.

The bureau previously signaled to NPR that those new test files may not be available until weeks after redistricting data is released, despite requests from data users, including members of the National Conference of State Legislatures.

"The Census Bureau recognizes that data users would benefit from access to the final set of [Privacy-Protected Microdata Files] for the redistricting data as soon as possible, and we are committed to releasing this information by September 30," the bureau said in a statement to NPR in June. "Given the delay in delivery of the redistricting data, however, we must prioritize production of the redistricting data so that states can commence already-delayed redistricting work."

Separate privacy protection decisions for other 2020 data sets are expected to be made later after gathering more public feedback. In May, the bureau's committee of outside scientific advisers recommended delaying the release of these data sets after redistricting data is out to give more time for testing the new privacy protections.

Why have the bureau's differential privacy plans been controversial?

Preliminary tests of the bureau's new privacy protections have left many data users worried that their ability to use 2020 census statistics could be severely limited, particularly data about small geographic areas and minority groups within communities that many governments rely on for planning.

Before announcing their finalized plans for redistricting data, bureau officials stressed, however, that their differential privacy plans were a work in progress as they gathered feedback from the public.

NATIONAL

Biden To Make Historic Census Director Pick With Latinx Statistician Rob Santos

Alabama filed a federal lawsuit in March to try to block the bureau from using differential privacy, which the state claims will make the data unusable for redrawing

voting maps. Sixteen states, most of which also have Republican-controlled legislatures, supported Alabama's claims in an amicus brief.

And more lawsuits over differential privacy may be coming later, including from civil rights groups that have been monitoring the bureau's test data to see whether the new protections make it harder to ensure fair representation of people of color during redistricting.

"At this point, it seems not at all clear that anything the bureau releases will eliminate the possibility that the Voting Rights Act and its enforcement could be adversely affected by differential privacy," Thomas Saenz, the president and general counsel of the Mexican American Legal Defense and Educational Fund who also serves on the bureau's National Advisory Committee on Racial, Ethnic, and Other Populations, told NPR before the bureau announced its finalized privacy settings for new redistricting data.

During a public meeting in May, that committee's chairperson — James Tucker, a voting rights attorney for the Native American Rights Fund — noted that many civil rights organizations "believe that the privacy concerns raised by the bureau are overstated."

"As bureau experts acknowledge, commercially available search engines that we use every day allow even the most unsophisticated data user to readily identify personally identifiable information about nearly any person they want," Tucker said, adding that there are "widespread concerns that the Census Bureau may be rushing in its embrace of differential privacy."

NATIONAL

After A Disrupted Census, Congress Tries Again To Extend Deadlines For Results

Bureau officials have previously countered similar arguments by noting that census data provides information that is often not well represented in commercial databases, including data about children, people's self-reported racial and ethnic identities, and same-sex couples who live in the same household.

"Given that the decennial census is a mandatory survey with universal coverage, the Census Bureau believes that it has both a legal and an ethical responsibility to use the strongest privacy protection technology available," the bureau said last year in a written response to committees of state demographers and data centers.

What happens if the courts block the bureau from using differential privacy?

Abowd, the bureau's chief scientist, warned in a court filing for Alabama's lawsuit that the release of 2020 census redistricting data — which is already late because of the coronavirus pandemic and the Trump administration's interference with the census schedule — could be further delayed by "multiple months" past August.

"This delay is unavoidable because the Census Bureau would need to develop and test new systems and software," Abowd added, later estimating that the work could last for at least six to seven months.

Editor's note: *Apple and Microsoft are among NPR's financial supporters.*

differential privacy 2020 census data census census bureau

More Stories From NPR