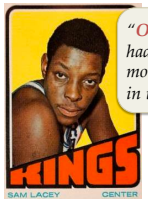
iCheck: Computationally Combatting "Lies, D—ned Lies, and Statistics"

You Wu† Brett Walenz† Peggy Li† Andrew Shim† Seokhyun Song† Emre Sonmez†
Pankaj K. Agarwal† Chengkai Li‡ Jun Yang† Cong Yu§
†*Duke University* ‡*Univ. of Texas at Arlington* §*Google Inc.*



"Only 10 players in NBA history had more points, more rebounds, and more assists per game than Sam Lacey in their career!"

"Jun had 3 SIGMOD papers and 3 ICDE papers in 2006; only seven other researchers have ever beaten this record"

Jim Marshall, a Democratic incumbent from Georgia, voted with Nancy Pelosi "almost 90 percent of the time"

Jim Marshall "is a long way from Nancy Pelosi," as he "voted the same as Republican leaders 65 percent of the time"

Rudy Giuliani, in a Republican presidential candidates' debate in 2007, claimed that "adoptions went up 65 to 70 percent" in the New York City "when he was the mayor"

How do we check these claims?

Opportunity: more and more structured data are available for checking

Crisis: Fact-checking is absurdly hard; traditional news media, in decline, can't keep up with investigative reporting needs

Leverage computing to "reverse-engineer" vague claims, assess quality beyond correctness, and come up with counterarguments

To check a claim, tweak how it manipulates data and see how conclusions differ

Challenge 1: *vagueness (by design?)*

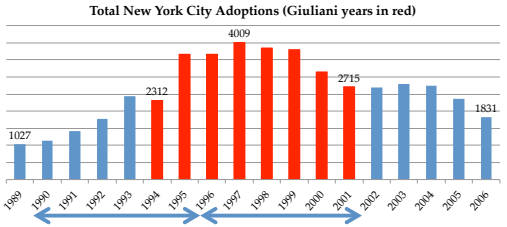
Adoptions went up 65 to 70 percent in NYC when Giuliani was the mayor Huh?

Total number of adoptions during 1996-2001 increased by 66.5% compared with 1990-1995 Hmm…

Giuliani was in office 1994-2001

Challenge 2: *beyond correctness*

Correct, but still misleading…



Challenge 3: *gimme the punchline*

Counterargument: Comparing the first and last years of Giuliani's tenure, adoptions increased by only 17%

Claim = a parameterized query q

- Has a specific setting p_0 of parameters
- Returns a specific answer $q(p_0)$

Perturb a claim: try different p from the parameter space $\mathscr{P}$ and compare $q(p)$ with $q(p_0)$

- Let $\mathrm{SR}(r; r_0)$ measure the strength of result r relative to r_0 (+ strengthens it; – weakens it)
- Let $\mathrm{SP}(p; p_0)$ measure the sensibility of parameter setting p given p_0 (pdf/pmf over $\mathscr{P}$— higher means more relevant to checking)

Query response surface (QRS): surface over $\mathscr{P}$ with height at $p \in \mathscr{P}$ given by $\mathrm{SR}(q(p); q(p_0))$

Claim quality measures = various aggregate properties of the QRS, e.g.:

Robustness: based on mean-squared relative strengths of perturbed claims, weighted by sensibility ☞ *e.g., apply to Giuliani's claim*

Uniqueness: probability of finding a perturbation that's weaker than the original claim ☞ *e.g., apply to Jun's claim*

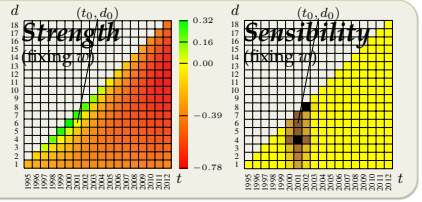
Reverse-engineering, finding counterarguments = optimization problems over the QRS, e.g.: above a sensibility threshold, find perturbations that match the original claim's result the best, or, resp., weaken it the most

Computational challenge: *efficient perturbation analysis of database queries*

- Exploit knowledge of queries
- Exploit properties of sensibility functions

Giuliani's claim, parameterized:

- Window aggregate comparison query; result: 66.5
- Three parameters: $w = 6$ (window length), $t = 2001$ (end of the second period), $d = 6$ (distance between two periods)



tinyurl.com/icheckuclaim

This is just the tip of an iceberg: join us in exploring endless possibilities for applying computing to fact-checking, journalism, and beyond!