Business Dynamics of Innovating Firms: Linking U.S. Patent Data with Administrative Data on Workers and Firms

Presented by Elisabeth Perlman

This represents the work of a great number of people including:

David Dreisigmeyer, Nathan Goldschlag, Matthew Graham, Stuart Graham, Cheryl Grim, Tariqul Islam, Marina Krylova, Alan Marco, Javier Miranda, Wei Ouyang, and Elisabeth Perlman

United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Additions to the Business Dynamic Statistics (BDS)

Several new projects:

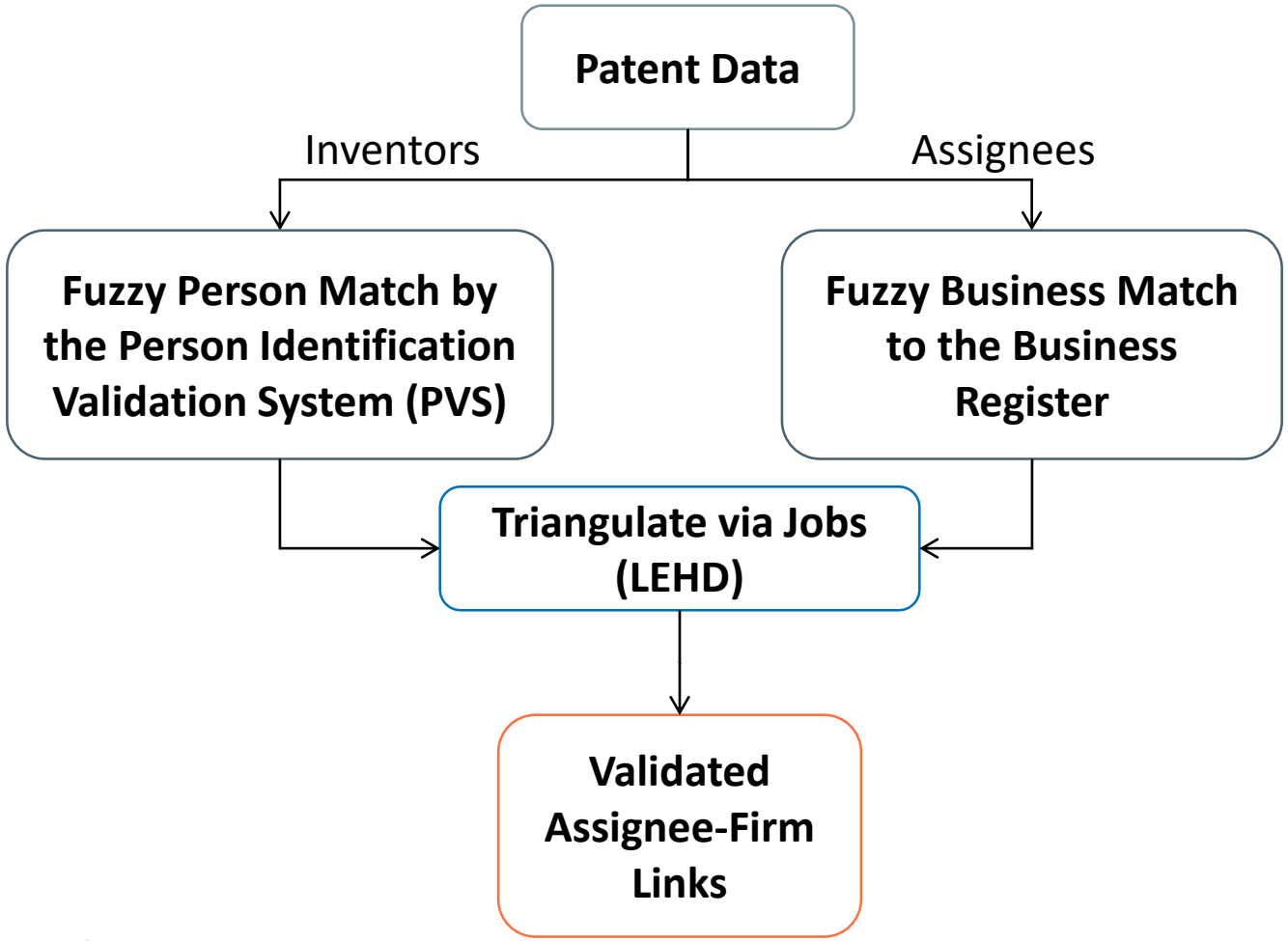
- Linking with administrative data for richer firm characteristics:
  - Export-Import data
  - Trademark data (public use)
  - Patent data (public use)
    - Objectives: Role of innovative firms in economic growth; Characteristics of innovative firms

- New privacy methods
  - New challenges for privacy when using public data
  - Implemental differential privacy (formally proven privacy)

Data inputs:
USPTO Data, Census Restricted-Use Data

- USPTO Custom Bibliographic Patent Data Extract (PTMT) (public use)
- USPTO Bulk Download Data (public use)
  - All granted patents between 2000 and 2015 (about 3 million patents)
- U.S. Census Bureau Business Register List (BR) and the Longitudinal Business Database (LBD)
  - Business list of all non-farm employer establishments with firm identifiers
- Longitudinal Employer Household Dynamics Employment History Files (LEHD-EHF)
  - List of employees covered by unemployment insurance provided by states.

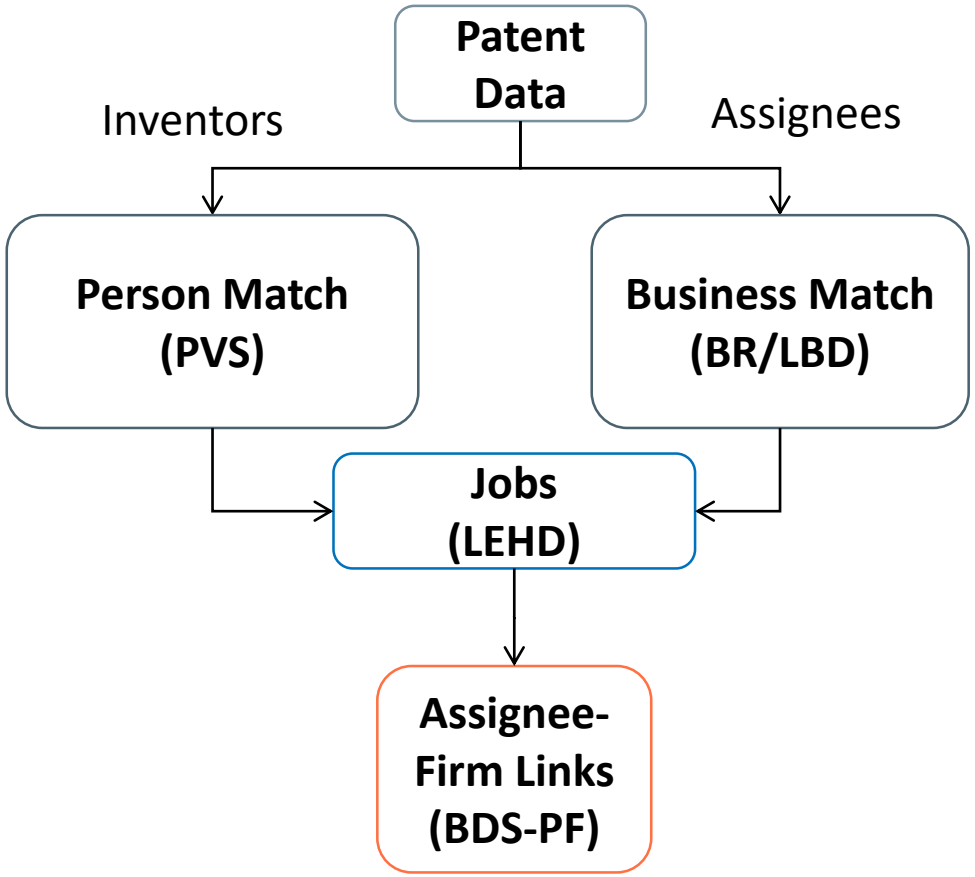
New Matching Strategy --- Triangulation



New Matching Strategy --- Triangulation

Triangulation of data allows:

- More precise matches
  - US assignee precision about 92%, foreign about 96%
- Higher match rates (closer to a true frame of patent holders)
  - US assignee match rate greater than 90%, foreign about 60%
- Validation of large number of matches
  - Fully triangulated matches have the highest precision



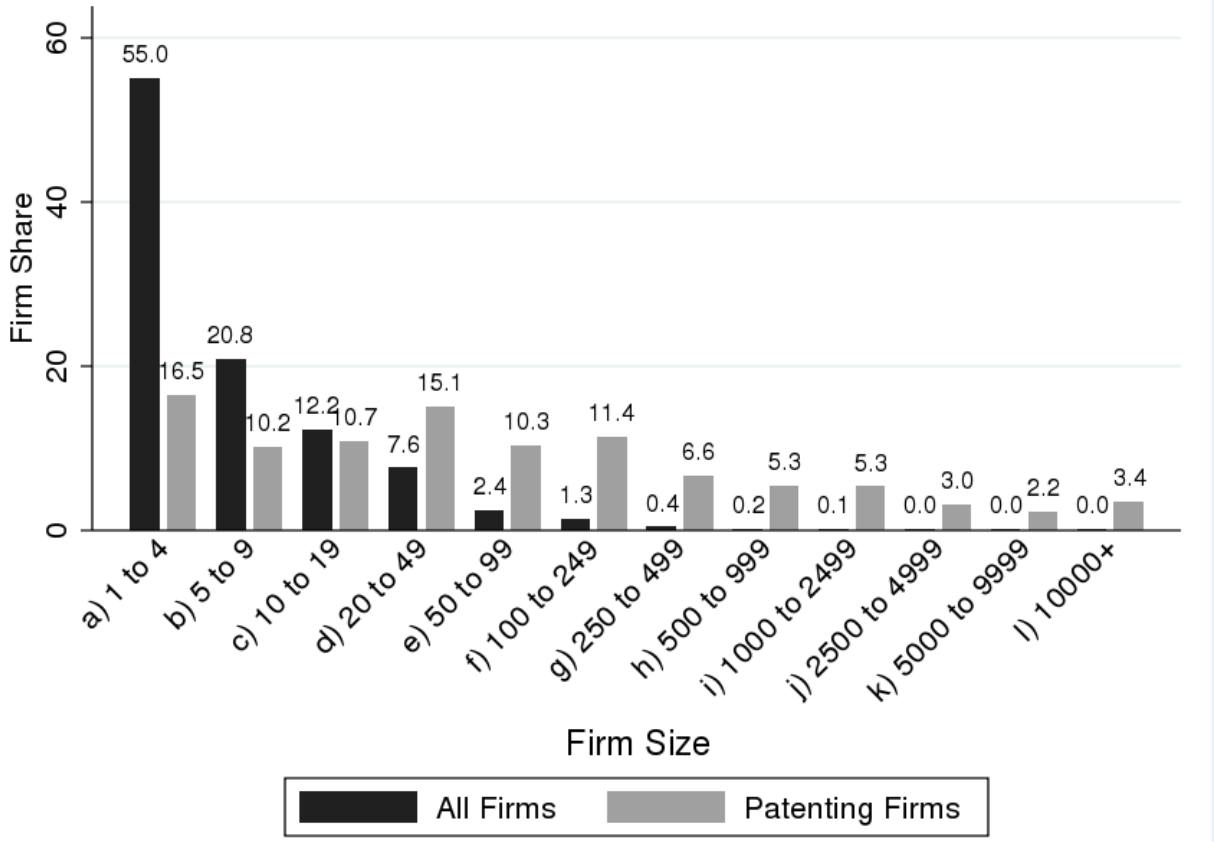
BDS Patenting Firms

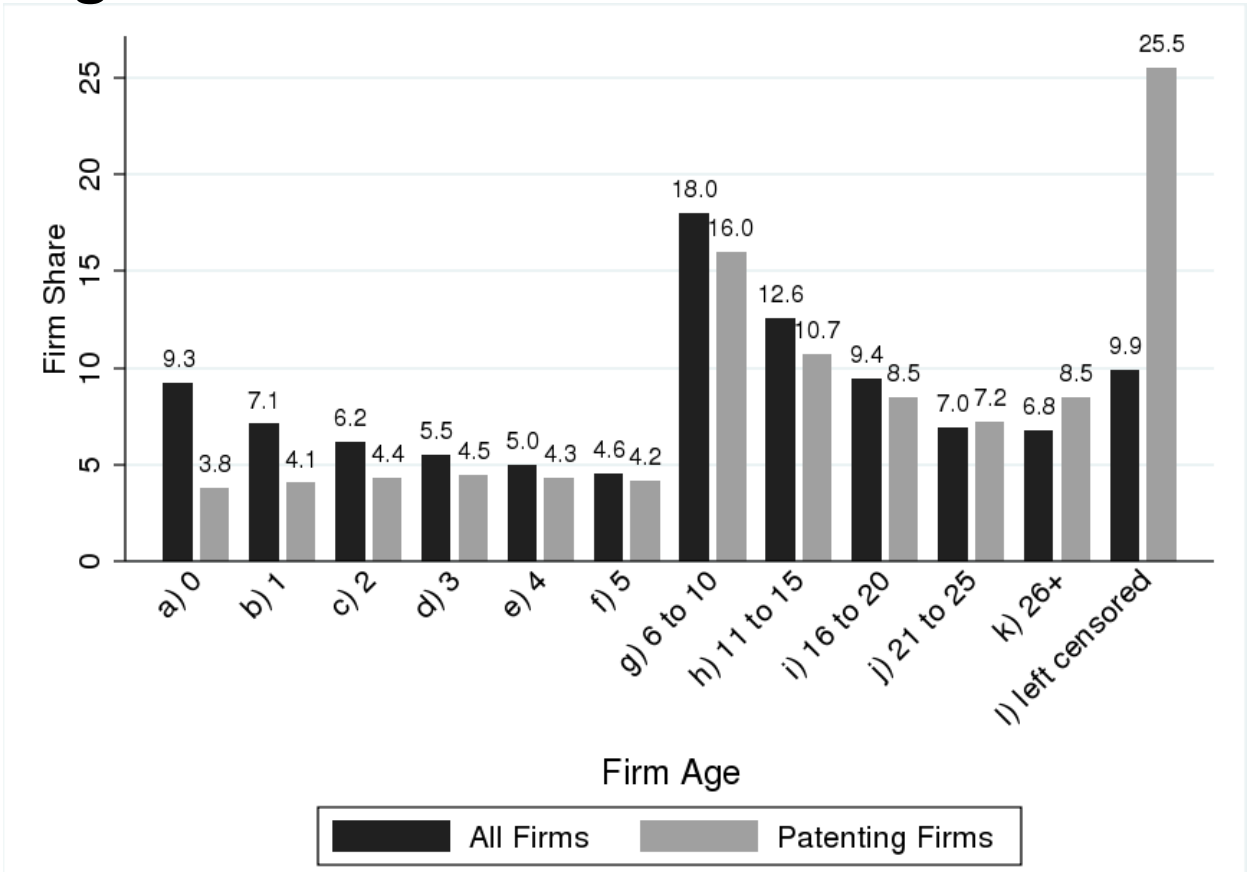
Data outputs

- Confidential longitudinal firm microdata: accessible through the U.S. Federal Statistical RDCs
  - Graham, S. J., C. Grim, T. Islam, A. C. Marco, and J. Miranda (Forthcoming) "Business dynamics of innovating firms: Linking US patents with administrative data on workers and firms," Journal of Economics and Management Strategy.
  - Dreisigmeyer, D., N. Goldschlag, M. Krylova, W. Ouyang, and E. Perlman (2018) "Building a Better Bridge: Improving Patent Assignee-Firm Links," CES Technical Notes Series, CES-TN-2018-01.
- Eventually public use tables, BDS of Patenting Firms (with new privacy methods)

Some Cross-Sectional Characteristics

Patenting firms are older and larger than firms as a whole.





What's Next in Addressing Privacy Concerns?

- We have been implementing rules of thumb to protect privacy (e.g. suppress any cell with fewer than N observations)
- How do you protect privacy when the list of entities (i.e. patenting firms) is known?
- Use methods that are formally proven

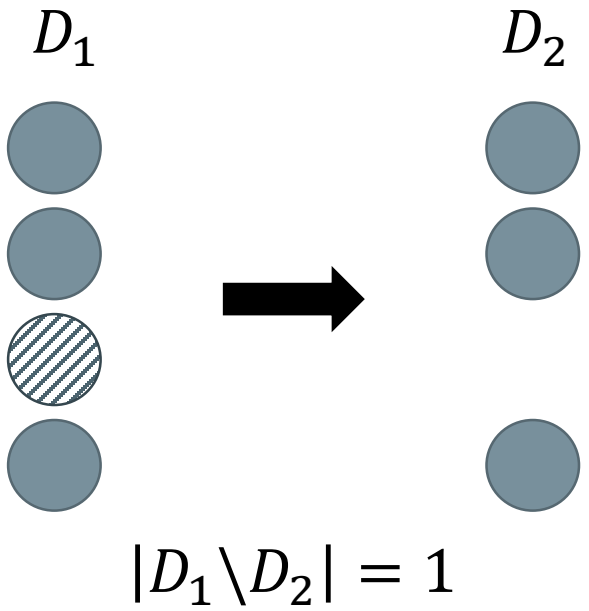
Differential Privacy (Formally Proven Privacy)

- This quantifies an upper-bound on what a person could learn from a given data release
- In doing so it provides data producers with a quantifiable trade-off between producing accurate statistics and protecting privacy
  - Improving one of these necessarily means decreasing the other.
- Thus data producers can work with a formal privacy-loss budget; an amount they can “spend” on a data release

Differential Privacy

- Privacy means that output of a Differential Privacy algorithm is insensitive to the addition/removal of one element

- Assume a worst case scenario attacker, knows every element but one
- After seeing the data how sure is the attacker about the element not in the data they already have?



Differential Privacy

- When dealing with counts of people, the addition or removal of one person changes counts by one

- But what if the statistic is biased on a firm's employment or people's wealth?
- How does one protect the existence of a very large firm?
  - Adding enough noise to keep the data from changing if that firm is inserted or removed will prevent statics about small firms from being meaningful

Differential Privacy

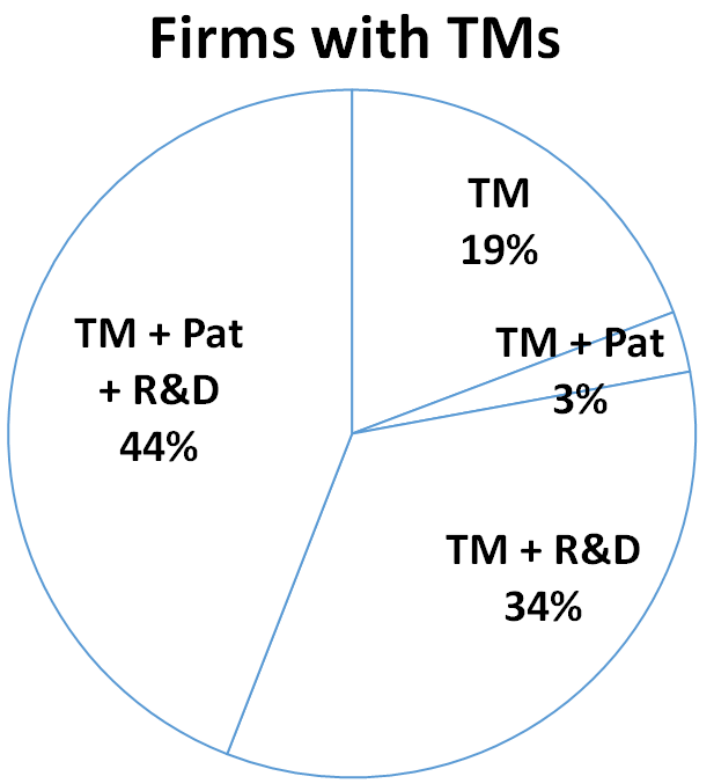
- One possible fix: worry about local sensitivity and infuse noise based the sensitivity on local areas of the data set (e.g. firm size)
  - E.g., firms with size 1-5 employees will receive noise that is scaled to a sensitivity of 5 and firms with 1000-5000 employees will receive noise that is scaled to a sensitivity of 5000
- However, if firms are aggraded geographically the existence of a single large employer in a small metro-area will not be protected, even if the particulars of that firm are
- Trade-offs are for data producers to decide

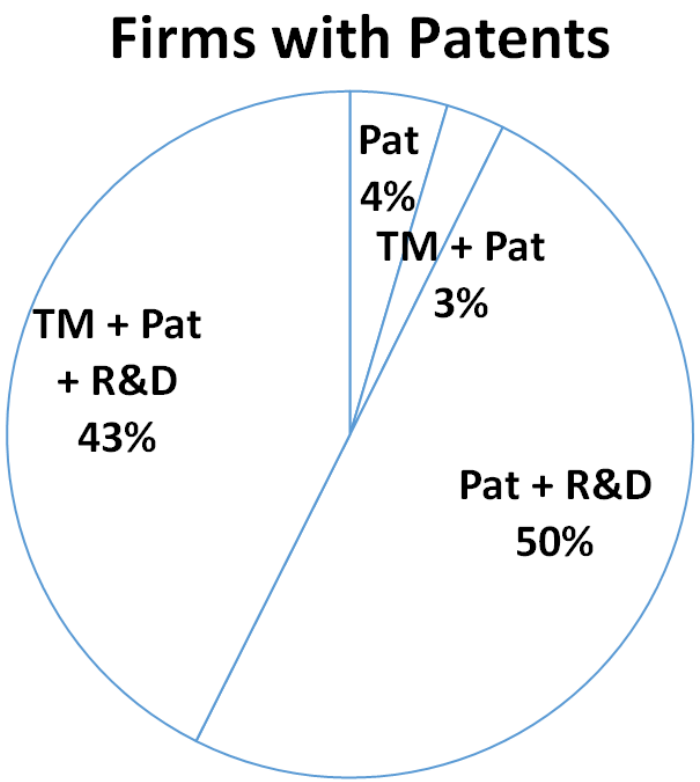
Questions?

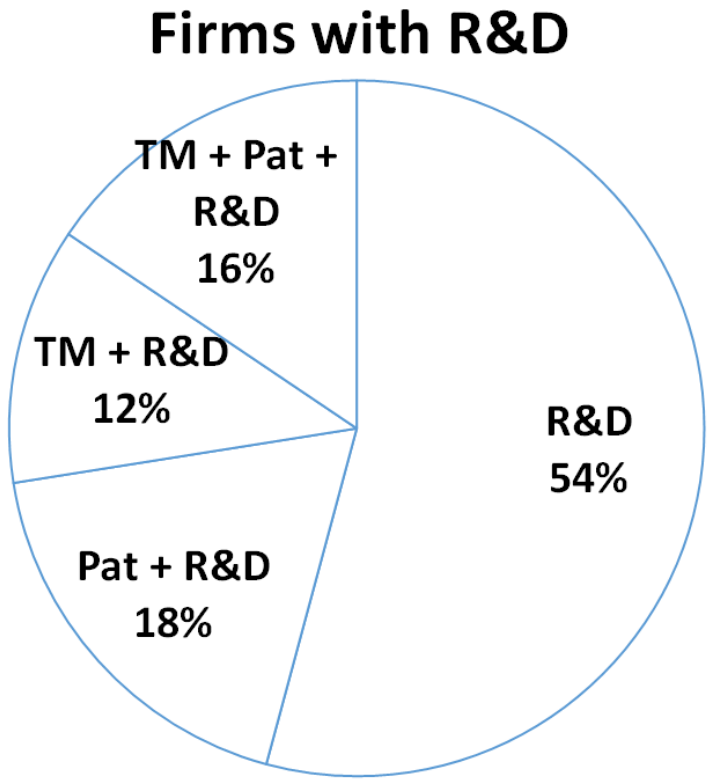
Differential Privacy

- One method for implementing differential privacy is adding Laplace-distributed noise.
  - This noise is not added to each entry in the dataset, but rather data aggregated to the smallest common cell that the data release requires
    - E.g. If we wish to release patenting firms by age and patenting firms by size as on slide 8, add noise to patenting firm cells broken by age X size
  - Noise must be scaled to the **sensitivity** of the query.
  - Sensitivity is generally the maximum amount a query can change if a single entity is added/removed from the underlying data.

Coincidence of Innovative Activities (BRDIS Sample)

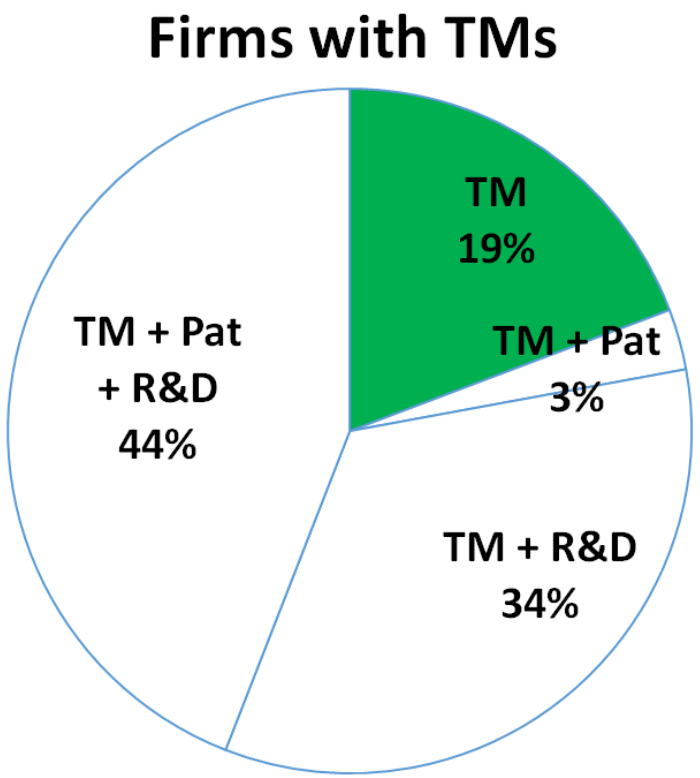


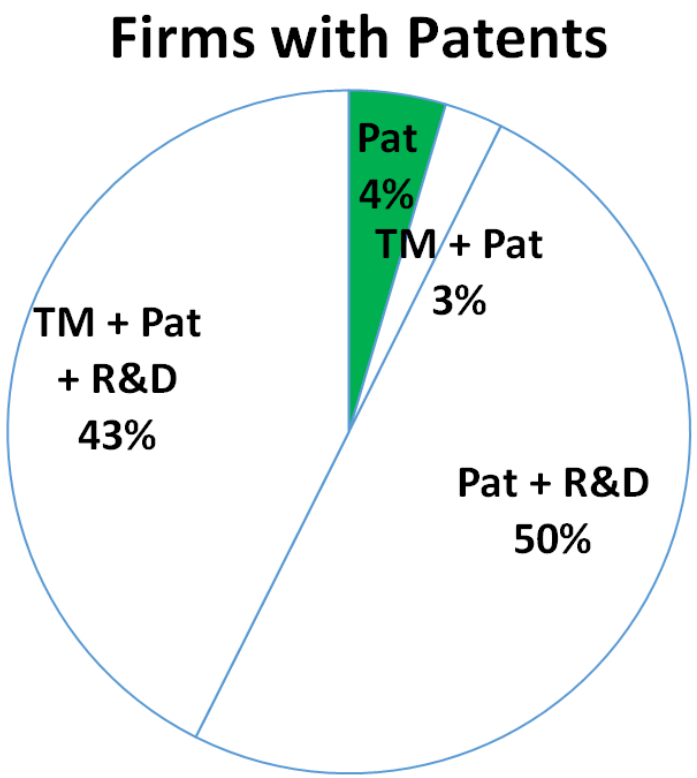


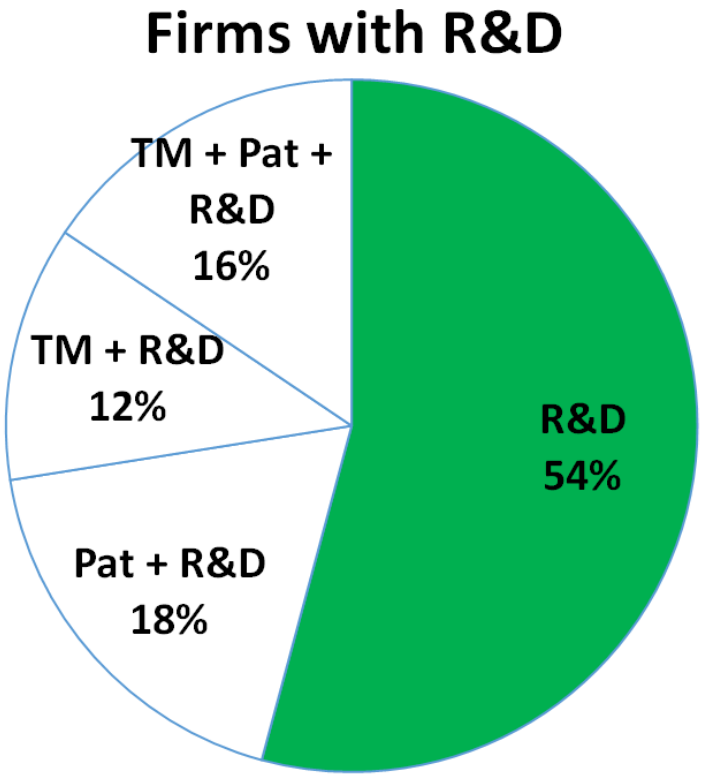


Coincidence of Innovative Activities (BRDIS Sample)

Only one activity





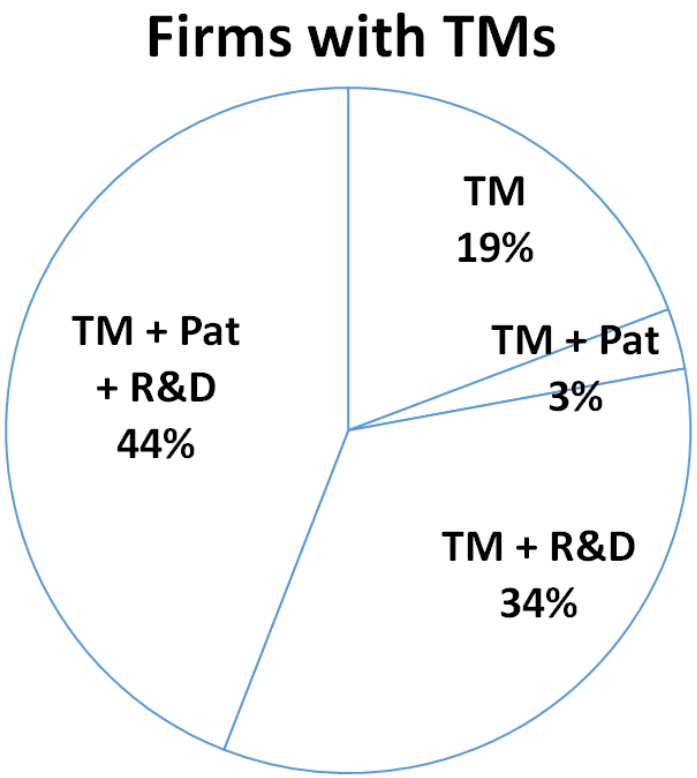


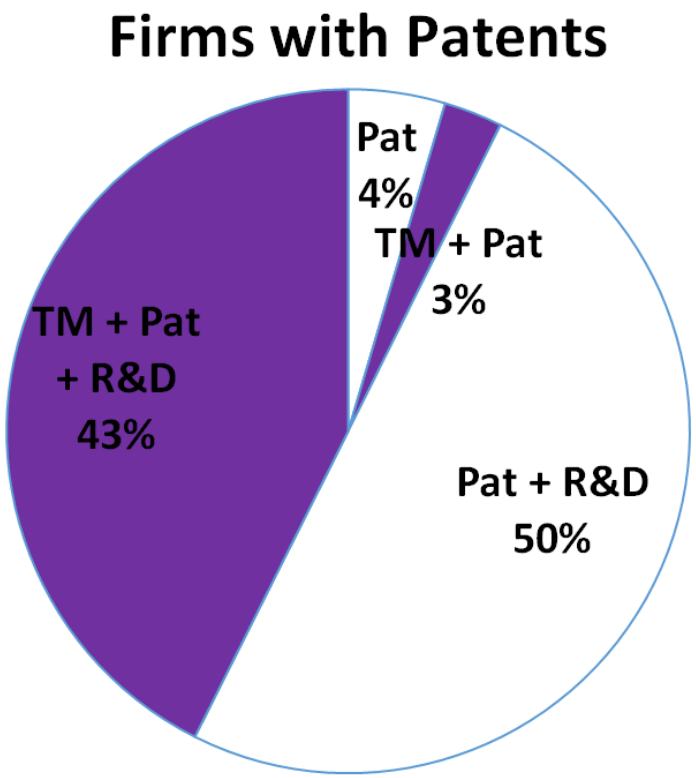
United States Census Bureau

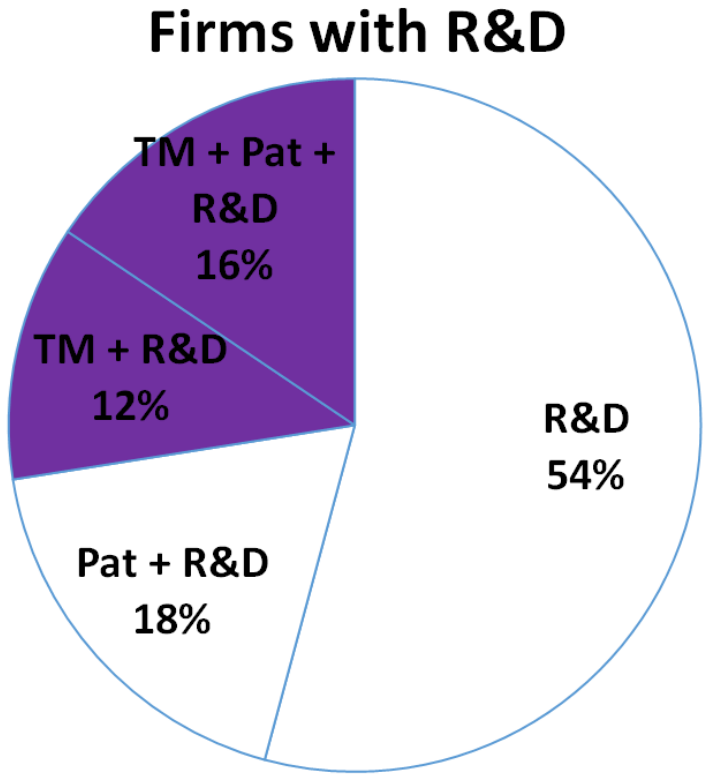
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Coincidence of Innovative Activities (BRDIS Sample)

Also trademarking





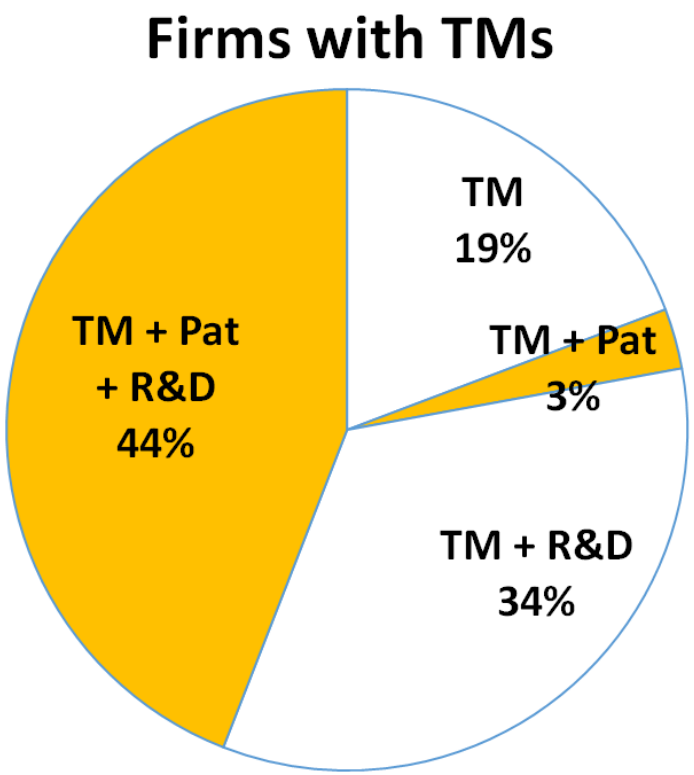


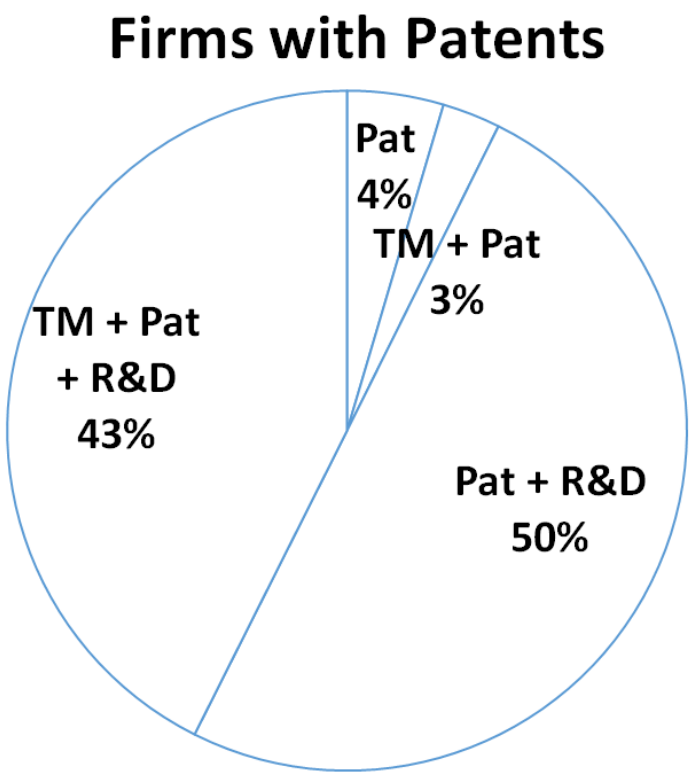
United States Census Bureau

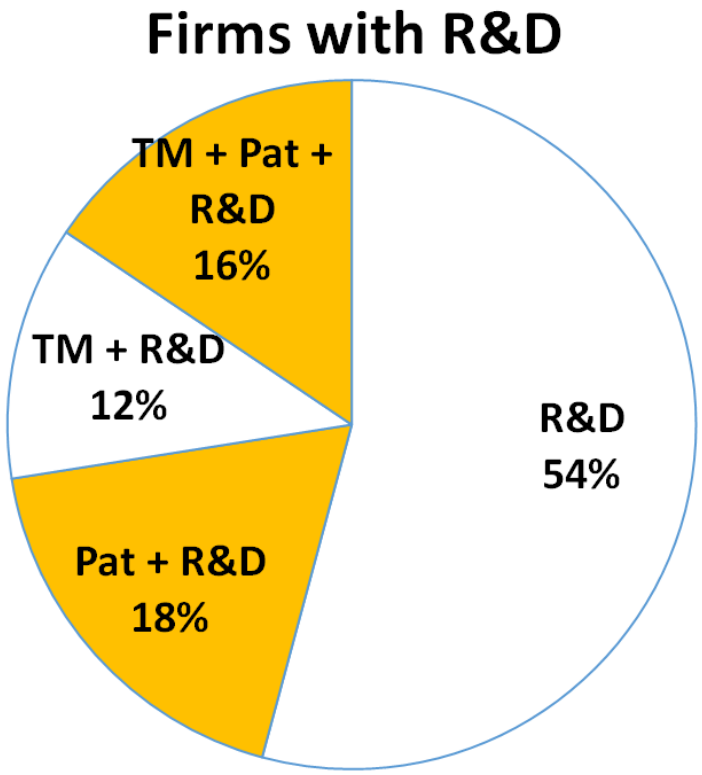
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Coincidence of Innovative Activities (BRDIS Sample)

Also patenting

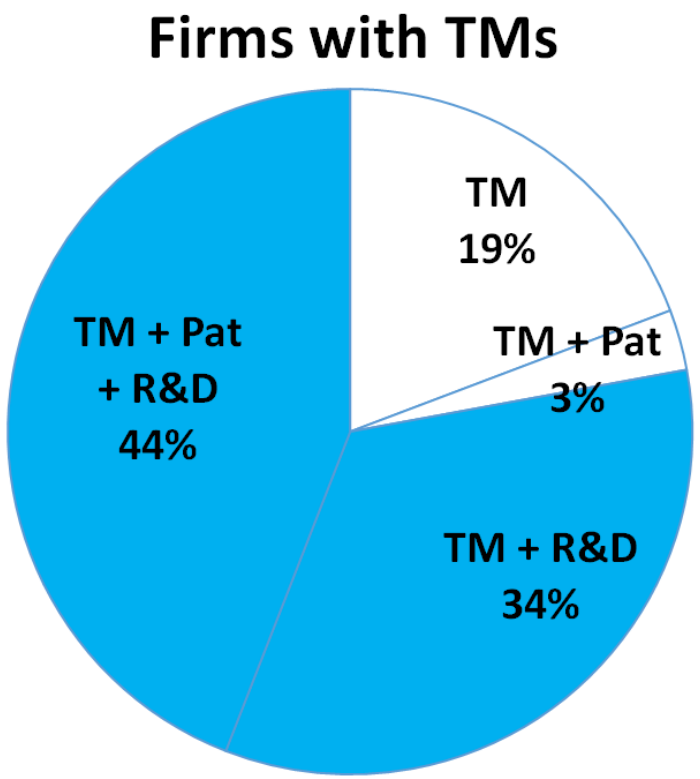


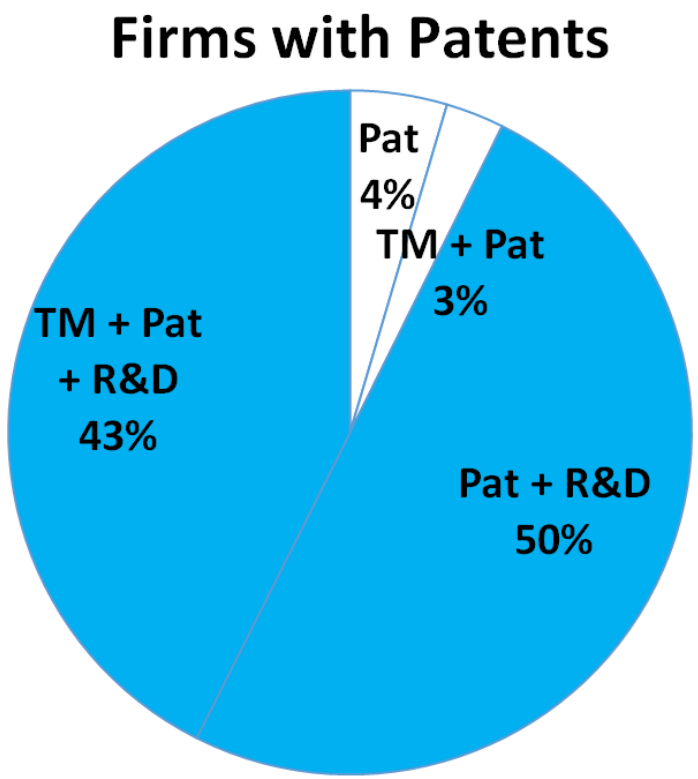


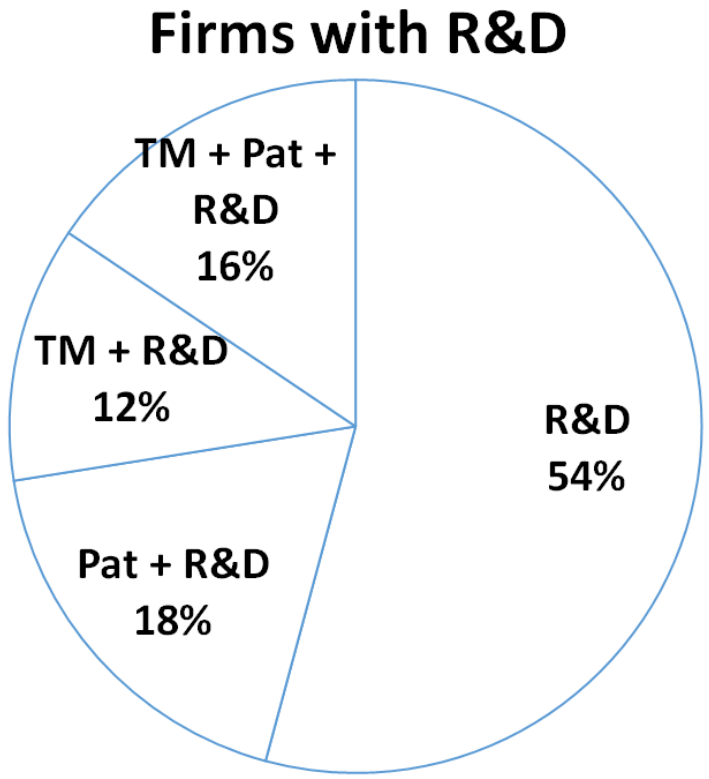


Coincidence of Innovative Activities (BRDIS Sample)

Also R&D







United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov