Online Access and Journal Citation Rates (some very preliminary results)

Mark McCabe
Georgia Institute of Technology

Christopher Snyder
Dartmouth College

Some Empirical Questions

- What is the impact of online access on journal citation rates?

- Are the benefits greater for newer or older content?

- Are the effects discipline-specific?

- Which online "channels" have the greatest impact?

- Does online access influence the geographic and institutional distribution of cited and citing authors?

Economic Context

- Scholarly Communication between Authors and Readers is mediated by (two-sided) journal platforms

- Similar to most information goods markets, online access has lowered content distribution costs

- This leads to "easier" access and possibly lower reader fees, new business models, changes in public policy, etc. The belief that citation rates are enhanced is an important assumption.

- More broadly, given the R&D/growth nexus, does online access have implications for rates of innovation, e.g. measured by patents?

The Natural Experiment

- Beginning in 1995 publishers and content aggregators began digitizing content and placing it online.

- However, as late as 2005 (the endpoint of our analysis) "backfiles" for many journals (and current content in some cases) remained offline.

- We exploit this heterogeneous chronology to explore the impact of online access.

- Past studies of this relationship (outside of the economics literature) report large effects.

- Common flaw: none of these efforts (adequately) control for potential selection problems.

The Data

- Online Chronologies for 300 "high-impact" journals in 3 disciplines, economics and business, general science and biology, history.

- For each journal we identified if and when each journal-year of content was placed online by their publishers, and content aggregators.

- Corresponding citation data at the *article* level was obtained from Thomson ISI for the period 1980- 2005.

- Specifically, for each year in this period we received all citations to Journal X, e.g. in 1990 we have all 1990 citations to any article published in the AER from 1886 to 1990, etc.

- The article level citation data is very detailed, and includes, for both citing and cited* articles, the authors' names, institutional affiliations, and locations, as well as article titles.

Some Descriptive Stats (for the econ/bus titles)

- 99 Journals, 3449 journal-year panel "groups, 58429 observations

	Mean	St Dev	Min	Max
First Publication Date	1947.93	26.56	1844	1988
Content Publication Date	1979.95	11.62	1956	2005
Citing Years	1994.89	7.13	1980	2005
Journal-Year Citations	37.02	60.00	0	771

Online Dates for 1980 Content, by channel

	Obs	Mean	St Dev	Min	Max
JSTOR	39	2000.82	2.64	1996	2005
Aggregator A	14	2003.00	1.57	2001	2005
Aggregator B	34	2002.18	1.47	2001	2005
Publisher Website	19	2001.95	0.23	2001	2002

82 Titles in the sample were published in 1980

Frequency versus Citations for all 99 titles, all years

journal-year citations

Frequency of Citations: 5,000 / 4,000 / 3,000 / 2,000 / 1,000 / 0

# of citations: 0, 200, 400, 600, 800, 1000
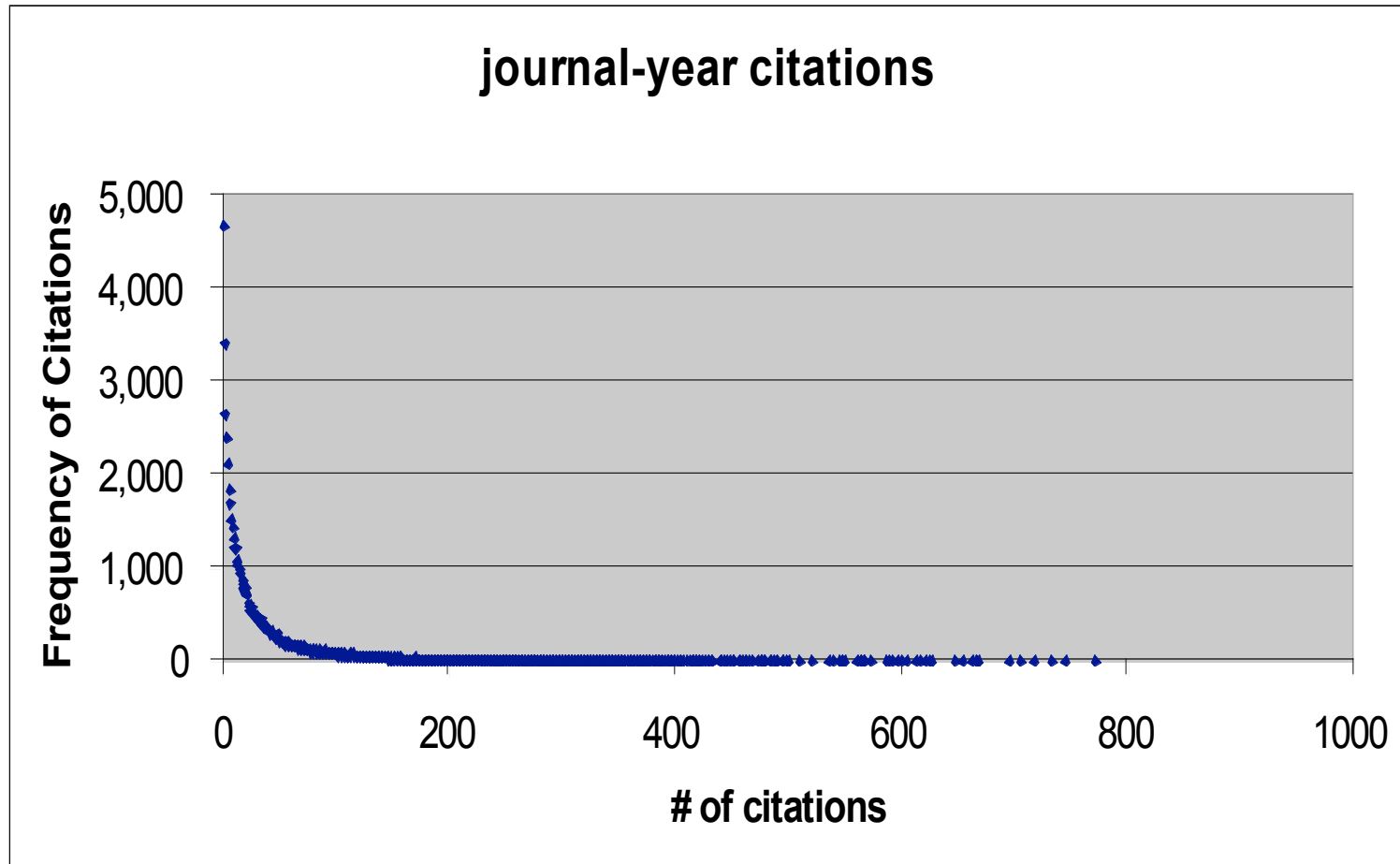
Methodology

- The non-negative integer nature of the citation data suggests a Poisson type process.

- The plot of frequency and citations implies that the data is overdispersed. So, a negative binomial specification is probably appropriate.

- The data is in panel form, and so fixed effects need to be considered

- We treat online access as an experimental "treatment". Dummy variables are specified to measure the impact of online access on citations.

Aggregate Online Effect, All Content

	IRR	[95% Conf. Interval]	
age	1.0777	1.0758	1.0796
age2	.99883	.99878	.99888
digit	1.0691	1.0503	1.0883

Dependent Variable: journal-year citations
IRR: incidence rate ratio

Aggregate Online Effect, 1956-1969 Content

	IRR	[95% Conf. Interval]	
age	.89899	.87923	.91919
age2	1.0009	1.0008	1.0011
digit	1.1230	1.0749	1.1732

Dependent Variable: journal-year citations
IRR: incidence rate ratio

Aggregate Online Effect by Year, 1956-1968 Content

	IRR	[95% Conf. Interval]	
age	.89705	.87720	.9173604
age2	1.0009	1.0008	1.001147
digit-1996	1.0788	.96298	1.208638
-1997	.99736	.88716	1.121239
-1998	1.0493	.94234	1.168587
-1999	1.0876	.97945	1.207895
-2000	1.1876	1.0680	1.320576
-2001	1.1236	.99722	1.26618
-2002	1.1579	.98196	1.365498
-2003	1.1067	.93967	1.303637
-2004	1.5488	1.2502	1.918871
-2005	2.0373	1.4749	2.814236

Dependent Variable: journal year citations
IRR: incidence rate ratio

Online Effect by Channel, 1956-1968 Content

	IRR	[95% Conf. Interval]	
age	.89629	.87646	.91656
age2	1.001	1.0008	1.0011
Channel A	1.0853	1.0271	1.1467
JSTOR	1.1309	1.0849	1.1789
Channel B	1.0918	1.0122	1.1776

Dependent Variable: journal-year citations
IRR: incidence rate ratio

Online Effect by Channel & Year, 1956-1968 Content

	IRR	[95% Conf. Interval]	
Channel A-2001	1.0561	.91758	1.2155
-2002	1.0499	.93512	1.1788
-2003	1.0073	.90184	1.1251
-2004	1.0760	.96414	1.2009
-2005	1.2422	1.1142	1.3851
Channel B-2001	.97792	.76712	1.2466
-2002	.95571	.75107	1.2161
-2003	.95343	.74669	1.2174
-2004	1.3142	1.1547	1.4956
-2005	1.0612	.94029	1.1978

Dependent Variable: journal-year citations
IRR: incidence rate ratio

Online Effect by Channel & Year, 1956-1968 Content, cont'd

	IRR	[95% Conf. Interval]	
JSTOR-1996	1.0797	.96388	1.2094
-1997	.99813	.88796	1.1219
-1998	1.0490	.94214	1.1681
-1999	1.0873	.97917	1.2074
-2000	1.1898	1.0701	1.3229
-2001	1.1234	1.0016	1.2600
-2002	1.0990	.95293	1.2676
-2003	1.1506	.99436	1.3315
-2004	1.3844	1.1939	1.6053
-2005	1.4363	1.1380	1.8127

Dependent Variable: journal-year citations
IRR: incidence rate ratio

Online Effect by Channel, 1995-2005 Content

	IRR	[95% Conf. Interval]	
age	2.0477	2.0021	2.09431
age2	.95541	.95397	.9568589
Channel A	1.1621	1.1203	1.205454
JSTOR	.92148	.89287	.9510184
Channel B	1.3156	1.2320	1.404963
Publisher Website	1.1558	1.1007	1.213738

Dependent Variable: journal-year citations
IRR: incidence rate ratio

Online Effect by Channel & Year, 1995-2005 Content

	IRR	[95% Conf. Interval]	
Channel A-1996	1.3428	1.0311	1.7486
-1997	1.2501	1.0710	1.4592
-1998	1.1496	1.0325	1.2801
-1999	1.1492	1.0534	1.2537
-2000	1.1330	1.0510	1.2214
-2001	1.1426	1.0729	1.2169
-2002	1.1898	1.1217	1.2620
-2003	1.1934	1.1287	1.2618
-2004	1.1782	1.1161	1.2438
-2005	1.1158	1.0547	1.1805

Dependent Variable: journal-year citations
IRR: incidence rate ratio

Online Effect by Channel & Year, 1995-2005 Content

	IRR	[95% Conf. Interval]	
Channel B-1996	1.0823	.89440	1.3097
-1997	1.0491	.92158	1.1944
-1998	1.0956	.98123	1.2234
-1999	1.1883	1.0775	1.3104
-2000	1.2758	1.1636	1.3988
-2001	1.3439	1.2315	1.4665
-2002	1.4175	1.3055	1.5391
-2003	1.3811	1.2756	1.4953
-2004	1.3981	1.2930	1.5117
-2005	1.3154	1.2150	1.4242

Dependent Variable: journal-year citations
IRR: incidence rate ratio

Online Effect by Channel & Year, 1995-2005 Content

	IRR	[95% Conf. Interval]	
JSTOR-1999	.62737	.40130	.98079
-2000	.85381	.72841	1.0008
-2001	.83420	.75707	.91919
-2002	.87470	.81988	.93319
-2003	.90523	.85386	.95970
-2004	.91924	.87051	.97071
-2005	.97944	.92572	1.0362

Dependent Variable – journal-year citations
IRR: incidence rate ratio