Appendix

Online material

An interactive map of the hyperlink data and the trade flow predictions between Local Authority Districts can be found in `url removed for anonymity`.

Out of sample R^2 for differnet sectors

Table 1: R-squared for t + 2 out of sample predictions per sector

sector	2002	2003	2004	2005	2006	2007	2008	2009	2010
s1	0.863	0.855	0.890	0.764	0.798	0.832	0.894	0.869	0.735
s2	0.885	0.861	0.864	0.866	0.856	0.907	0.901	0.892	0.797
s3	0.892	0.823	0.815	0.795	0.846	0.857	0.904	0.885	0.772
s4	0.739	0.805	0.790	0.871	0.876	0.876	0.880	0.788	0.802
s5	0.615	0.782	0.717	0.738	0.784	0.752	0.725	0.580	0.587
s6	0.917	0.837	0.902	0.911	0.882	0.926	0.947	0.910	0.893
s8	0.970	0.951	0.936	0.905	0.914	0.907	0.921	0.936	0.909
s9	0.976	0.968	0.974	0.954	0.972	0.973	0.983	0.980	0.890
s10	0.944	0.953	0.926	0.903	0.930	0.927	0.951	0.934	0.640
s11	0.664	0.695	0.183	0.420	0.398	0.602	0.725	0.618	0.204
s12	0.935	0.845	0.926	0.896	0.905	0.932	0.955	0.940	0.524
s13	0.748	0.764	0.874	0.838	0.869	0.835	0.877	0.882	0.478
s14	0.721	0.668	0.817	0.811	0.795	0.711	0.805	0.914	0.332
s15	0.815	0.773	0.947	0.911	0.910	0.822	0.870	0.934	0.534

s1: Agriculture, s2: Mining, s3: Food, s4: Textiles, s5: Chemicals, s6: Equipment,
s8: Manufacturing; s9: Construction, s10: Distribution, s11: Hospitality,
s12: Transport, s13: Financial, s14: Real Estate, s15: Non-Market Services

Lasso regressions

Table 2: LASSO regressions: accuracy metrics in unseen data from t + 2

year	RMSE	Rsquared	MAE
2002	3416.15	0.64	1233.66
2003	16642.84	0.24	2172.69
2004	2887.28	0.64	1090.96
2005	4441.22	0.43	1006.57
2006	4418.99	0.36	1153.61
2007	4781.95	0.38	1222.12
2008	6178.04	0.26	1262.79
2009	3918.96	0.42	1259.55
2010	4445.62	0.3	1150.39

Data wrangling process

This section describes how geographic regions were added to the host-linkage dataset provided by JISC UK Web Domain Dataset[1]. The archived web data are from 2000-2010. The process begins combining the

[1] https://data.webarchive.org.uk/opendata/ukwa.ds.2/geo/

host-host links to a file containing unique postcodes for each host ending in co.uk. An example is provided below.

Table 3: Host-linkage file

year	origin	destination	links
2000	btclickbus.excite.co.uk	greenwich2000.co.uk	1
2000	btclickfree.excite.co.uk	www.rockvillecenter.com	2
2000	adapthorpe.com	www.adapthorpe.com	1
2000	btclickfam.excite.co.uk	conciergedesk.co.uk	1
2000	formby.wiganmbc.gov.uk	www.charitynet.org	1

Table 4: Websites with unique postcodes

URL	postcode	year	host	domain
20000609075945/http://altberg.co.uk:80/military_boots.htm	DL10 4XB	2000	altberg.co.uk	altberg
20001003045622/http://www.guest-house.demon.co.uk:80/	CB2 1AA	2000	www.guest-house.demon.co.uk	demon
20000917204128/http://www.millenniumit.co.uk:80/CV.htm	E3 5AN	2000	www.millenniumit.co.uk	millenniumit
20000312143711/http://www.nova-tech.co.uk:80/page2.html	PR9 9DZ	2000	www.nova-tech.co.uk	nova-tech
20000914061255/http://www.aleontap.co.uk:80/weblinks/	WR6 6DH	2000	www.aleontap.co.uk	aleontap

The two data frames were joined by matching the variable `domain`. If an origin or destination was found in in the postcode data, it was added to the file. Host-links without a postcode were dropped. This leaves with host, domain and postcode for origins and destinations and the number of links between. This is shown below.

Table 5: Combined host and postcode data

origin.host	orig.domain	origin.pc	dest.host	dest.domain	dest.pc	links
24carat.co.uk	24carat	FY4 1RJ	24carat.co.uk	24carat	FY4 1RJ	9201
www.lifestyle.co.uk	lifestyle	WC1A 2AE	www.lupine.demon.co.uk	demon	KT17 2HB	3
www.barcodes-for-access.beechman-online.co.uk	beechman-online	BR1 1PD	www.ao-plotters.beechman-online.co.uk	beechman-online	BR1 1PD	3
www.bringfrd.demon.co.uk	demon	NN6 6HB	www.bringfrd.demon.co.uk	demon	NN6 6HB	28

The next step was to remove websites that linked to themselves (e.g. the first row above). These data were not of interest as we are looking for links between different websites. Therefore, if origin host and destination host were the same, they were dropped. We now have host-to-host links with the associated unique postcodes and the number of links.

The next step was to aggregate to the NUTS2 regions. This was done by using a postcode to NUTS2 (2010 version) lookup file combined with the above created data. The data was then aggregated summing all data with the same origin NUTS and destination NUTS codes. We are then left with our NUTS2-to-NUTS2 links. The same process was done for every year 2000-2010.

origin.host	orig.domain	origin.pc	dest.host
24carat.co.uk	24carat	FY4 1RJ	24carat.co.uk
www.lifestyle.co.uk	lifestyle	WC1A 2AE	www.lupine.demon.co.uk
www.barcodes-for-access.beechman-online.co.uk	beechman-online	BR1 1PD	www.ao-plotters.beechman-online.co.u
www.bringfrd.demon.co.uk	demon	NN6 6HB	www.bringfrd.demon.co.uk

Table 6: NUTS2 level data

origin	destination	weight
UKC1	UKC1	90
UKC1	UKC2	1
UKC1	UKE2	1
UKC1	UKH3	1
UKC1	UKI1	3