

E-mail decay rates among corresponding authors in MEDLINE

The ability to communicate with and request materials from authors is being eroded by the expiration of e-mail addresses

Jonathan D. Wren, Joe E. Grissom & Tyrrell Conway

One of the early features of ARPANET, the predecessor to the modern-day Internet, was the ability to communicate via electronic mail (e-mail) with other users; the first e-mail was reportedly sent in 1971 (Hardy, 1996). E-mail quickly gained popularity among computer scientists, but spread more slowly through the biomedical research community. The first article in MEDLINE to mention 'electronic mail' appeared in 1983 (Cross, 1983). After another decade—about the same time that graphical methods of Internet navigation, namely web browsers, became popular—many MEDLINE-listed journals began to include electronic contact information in articles, which led to a rapid growth in the number of publications with corresponding author e-mail addresses (CAEs; Fig 1).

The potential benefits of e-mail as a means of correspondence in biomedical research was noted even before the start of this trend (Pallen, 1995). Now, e-mail is not only commonplace in biomedical research, it is the norm: authors, collaborators, readers, editors and reviewers for scientific manuscripts all rely on e-mail. We generally do not think about how much it has transformed the way we communicate with our colleagues until our e-mail servers are down. Although its ubiquitous nature renders e-mail a 'transparent technology', it is hard to ignore the advantages it holds over previously dominant means of scientific correspondence: postal and voice communications. The difference in cost and effort between sending one and many e-mails is minimal, unlike other media. E-mail

communication may be slower than voice, but it is virtually free—assuming computer equipment and Internet connections are in place—even across great geographical distances. E-mail also permits the exchange of several electronic media, such as text files, images and movies, and allows users to keep an archive of their correspondence without sacrificing office space to growing piles of paper. Owing to the reduced effort and cost compared with postal mail or telephone, and the fact that e-mail communication is generally viewed as less formal (Pallen, 1995), e-mail has facilitated more frequent exchanges among scientists, even if some researchers may receive more e-mails than they can handle (Singarella *et al*, 1993).

Now, e-mail is not only commonplace in biomedical research, it is the norm...

However, e-mail accounts are certain to have a finite existence, just as people do. It is doubtful that e-mail accounts even last as long as a human lifespan, as many people are prone to move or change jobs. Several studies using e-mail surveys have noted that a significant fraction of their recipient addresses were invalid (Knauper *et al*, 2004; Nguyen & Murphy, 2001; Treadwell *et al*, 1999). What is not known, however, is the rate at which e-mail accounts become invalid. This will probably vary by individual circumstance and profession, but increased knowledge of the e-mail decay rate can help with planning efforts. For

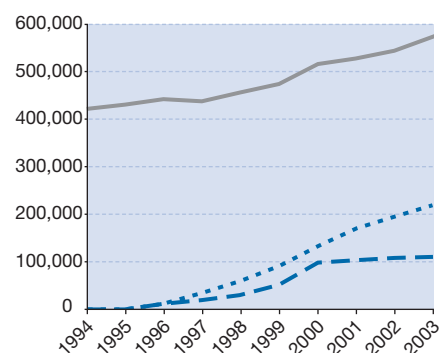


Fig 1 | Number of e-mail addresses published in MEDLINE per year compared with the number of articles per year (solid line). Also shown is the number of unique e-mail addresses published (dashed line), which is growing more slowly than the total number (dotted line).

example, conducting surveys by e-mail has the advantage that the response rate is higher and people tend to be more candid in their responses (Oppermann, 1995; Thach, 1995). Thus, researchers contacting study participants over time using e-mail surveys need a better understanding of how many e-mails are likely to be active at the end of a given time period.

Consider also the scientific endeavour, in which the end product is data and research results. All the products of a research project are rarely contained solely within a single published paper, and a single paper will rarely address all questions of interest. For researchers, patients and other consumers of biomedical research information, it is

therefore important to be able to contact the authors for clarifications, requests for materials, supplementary data or article reprints. E-mail is the most convenient way of contacting people, and anything that can be done to alleviate e-mail expiration will improve research-related communication.

The corresponding author (CA) holds a unique position on any paper, as the person designated to answer questions and handle correspondence for all other authors. Presumably, the CA had the broadest control over the published research project and retains the most knowledge, such that they can fulfil requests for experimental details, materials, software, databases, data sets or other information. If the CA does not have information on the published research, then they should know which of the co-authors should be approached. Yet, the selection of a CA may also vary by culture or nationality—a student may have been closely involved in the work and is thus best equipped to answer questions, but the senior author will probably remain at the same institution and e-mail address for longer.

E-mail is the most convenient way of contacting people, and anything that can be done to alleviate e-mail decay will improve research-related communication

When a CA leaves an institution or changes e-mail address, this clearly has an impact on future correspondence for all studies published before the move. In cases where an author's name is unique, the Internet can help interested parties to find them. This assumes that the CA or the institution is reasonably diligent in publishing contact information on the web, where a search engine such as Google or Yahoo will eventually index it. However, it should be noted that search-engine indexing is incomplete and sometimes does not include even half the materials available on a root website (Wren, 2005). And in cases where a CA's name is more common (for example, 'J Smith', 'W Wang', 'Z Chen'), they can be much more difficult to locate.

Unlike names, which may be shared by many individuals (Torvik *et al*, 2003) or change after marriage, e-mail addresses must be unique and thus permit us to study

collaboration networks more accurately than author names. Previous studies in locating individuals via an e-mail network suggest that it may be necessary to contact from one to seven individuals, at a minimum, to obtain the new e-mail address of a CA who has moved (Dodds *et al*, 2003). The most important issue is not whether the CA can be located, but rather how much time and effort it requires and whether simple steps can be taken to increase the continuity of e-mail contact.

We are increasingly using Internet resources in biomedical publications, such as e-mail addresses for correspondence and uniform resource locators (URLs) to provide online supplemental information and computational resources. Several studies have shown that URLs cited in research papers are far from permanent and that they disappear in a time-dependent manner after publication (Dellavalle *et al*, 2003; Lawrence *et al*, 2001; Spinellis, 2003; Wren, 2004). Compared to previous studies on URLs within MEDLINE (Wren, 2004), we find here that e-mail addresses are more prevalent by a couple of orders of magnitude. It is reasonable to expect e-mail accounts to expire in a similar manner. However, e-mail addresses differ from URLs in that some e-mail addresses are intended to 'decay'—for example, when users change their e-mail address after changing jobs or if they feel overwhelmed by unsolicited, 'spam' e-mails. By better understanding when and what types of e-mail addresses become invalid, we can infer why and make informed policy decisions. We can also get an idea of author mobility rates by analysing e-mail account availability by e-mail domain—for example, on the country or university level—more easily and cheaply than by postal or telephone inquiries. E-mail decay has not been studied previously, in part because checking e-mail availability without 'spamming' people is not as technically straightforward as checking a URL, which can be done by anyone with a web browser.

We first extracted all electronically available e-mail addresses published as contact information for CAs as of 15 June 2004, from MEDLINE and found a total of 1,017,205 e-mail addresses, 580,548 of which were unique. The US National Library of Medicine (Bethesda, MD, USA) graciously provided MEDLINE records in XML format, which enabled us to

locate and extract the CAE as identified by the presence of an '@' symbol in a contiguous block of text from the 'affiliation' field. CAEs were deposited into a database along with the dates when they were first and last linked to a publication.

E-mail accounts were queried for their status using Internet communication protocols, i.e. Simple Mail Transfer Protocol (SMTP). Although not all e-mail addresses can be checked using this method, it was preferable to sending out approximately 580,000 unsolicited e-mails. Perhaps most importantly—in addition to the fact that our institution has policies against large-scale unsolicited e-mailing—we personally detest receiving spam and would prefer not to engage in it. Also, there is no easy way to send, receive and analyse 580,000 e-mails using most commercial software, and it is necessary to repeat the analysis to ensure that invalid domains/e-mails were not failing to respond due to temporary problems. Finally, this method is preferable for future study reproducibility, as well as for the re-examination of e-mail decay rates to monitor changes in trends.

The steps we used to process and query the e-mail addresses are shown in Fig 2. As some domain name servers (DNSs) always return a valid response when an e-mail account is queried, we used a randomly generated bogus 8-character username to check each new domain to see if it fell into this 'accept all' category. If the DNS approved the bogus username as valid, we considered any e-mail address within that domain as unable to be queried.

Each e-mail address in the database was queried on three separate occasions to ensure that results were reproducible and to reduce error rates caused by temporary communication problems. Such communication errors fell into several categories: "Cannot connect SMTP" errors indicated that the domain was found and had a valid mail exchange (MX) record, but for some reason the internet protocol (IP) address did not respond. "DNS failure: NOERROR" indicated that the domain was found but without an associated IP address. "DNS failure: SERVFAIL" indicated that the server encountered an internal failure while processing the query request. "DNS timeout" appears if the e-mail server does not respond within 60 seconds, similar to a "SMTP timeout". We were reluctant to group communication-based error messages into the unavailable e-mail account

category, but for summary purposes chose to count them as such because communication errors were relatively consistent between the three runs. Of e-mails with "Cannot connect SMTP" errors on the third run, 76% returned errors on the first and second run as well, as did 82% of the "DNS failure: NOERROR" errors, 93% of the "DNS failure: SERVFAIL" errors, 69% of the "DNS timeout" errors and 23% of the "SMTP timeout" errors. Overall, 76% of communication errors were persistent across all three runs, suggesting that most of these failures were not due to temporary problems. Excluding communication errors from the analysis did not affect the time-dependent trend in CAE availability.

The most important issue is not whether the CA can be located, but rather how much time and effort it requires...

Invalid e-mail accounts return an error code that falls into one of the following five categories: "Invalid domain name syntax" or "Invalid username syntax", which are relatively self-explanatory; "DNS failure: NXDOMAIN", which indicates a nonexistent domain or that the server has no record of any type for the requested name, including generic top-level domain errors; and "no such user", which explicitly states that the account in question does not exist.

It took approximately one month to check the status of our set of published CAEs, with most of this time spent waiting for the 60-second time-out responses. The first run ended on 19 July 2004, the second on 3 September 2004 and the third on 17 November 2004. The results were consistent across each run, the details of which are provided as supplementary information on our website. The statistics reported hereafter are from the third and final run.

The oldest e-mail address we found within MEDLINE was published in 1986, with only 33 others found until 1995. Thereafter, the number of CAEs published in MEDLINE grew rapidly compared with the growth of MEDLINE overall (Fig 1). But Fig 1 also shows that the number of unique CAEs in MEDLINE has increased at a much slower rate over the past few years than the total number of CAEs published. This indicates that much of the research being published comes from existing CAs

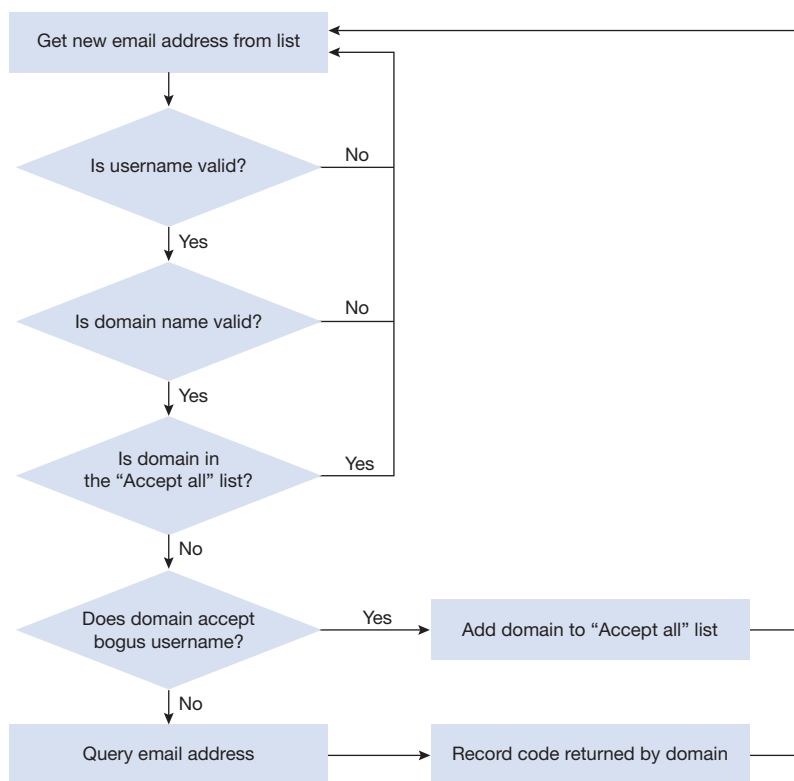


Fig 2 | Method for processing and querying e-mail addresses. Scripts to check the availability of e-mail addresses were written in Perl and run on a Sun Ultra-10 400 MHz Sparc workstation running Solaris 9 with a 100 Mbps Full-Duplex Ethernet adaptor.

rather than newcomers and suggests that the average CAE is linked to multiple publications. Mark Newman (University of Michigan, Ann Arbor, MI, USA) recently published a viewpoint that there exists a relatively modest number of "leaders" or influential people in biomedical research with the rest being "followers" or peripheral actors (Newman, 2004), which our data seem to support.

The results of the study and the error codes associated with invalid e-mail addresses from the time they first appeared in MEDLINE are summarized in Table 1. Approximately 44% of the e-mail addresses in the database belonged to an 'accept all' domain, meaning that reliable information could only be obtained for 255,031 out of 580,548 e-mail addresses. It is not known whether e-mails in 'accept all' domains deactivate at a different rate, although there is no apparent reason to suspect that there is a difference. The number of e-mails associated with an 'accept all' domain has been growing but appears to be tapering off in recent years (Fig 3).

The percentage of e-mail addresses that returned an error code, by year of their first appearance in MEDLINE is shown in Fig 4. As expected, the graph shows a gradual decline in e-mail account availability over time. However, it was surprising to find that 24%—almost one in four—e-mail accounts become invalid within a year of being published. Domain failure accounts for an average of 32.7% of these invalid e-mail addresses, although the shape of the curve indicates that this was a bigger problem before 2000 and may have diminished since then. As domains are the central communication hubs by which e-mail accounts are handled, the loss of a domain means that all e-mail addresses associated with it are lost as well. To provide a perspective on how domain loss affects e-mail address loss, Fig 5A shows the number of e-mails attached to each domain within MEDLINE, and Fig 5B shows the number of e-mails attached to domains that the DNS indicated were non-existent. Both distributions appear to follow an inverse power-law distribution, which suggests that a few domain

Table 1 | Summary of e-mail account status codes for all e-mail addresses that permit a query (e-mail addresses that do not belong to a domain that will declare any e-mail address to be valid)

E-mail account status	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	All years	Percentage
OK	–	82	1,263	2,667	4,351	10,370	24,316	26,363	28,755	30,317	12,425	140,909	55.3
Communications failure*	2	81	1,104	1,832	2,484	3,260	4,593	3,608	2,864	2,310	846	22,984	9.0
Bad domain name syntax	1	5	95	149	194	219	250	265	205	226	90	1,699	0.7
Bad username syntax	–	3	23	45	59	58	65	75	70	65	44	507	0.2
Non-existent domain	3	216	2,524	4,247	4,852	5,261	5,802	4,753	4,313	3,998	1,373	37,332	14.6
No such user	3	115	1,841	3,515	4,961	7,278	10,793	9,255	7,111	5,156	1,562	51,590	20.2
Invalid/inactive accounts	9	420	5,587	9,788	12,550	16,076	21,503	17,956	14,563	11,755	3,915	114,122	44.7
Total e-mails	9	502	6,850	12,455	16,901	26,446	45,819	44,319	43,318	42,072	16,340	255,031	100.0
Percentage failed	100	84	82	79	74	61	47	41	34	28	24	44.7	–

*This is a general category that indicates communication with the e-mail server could not be established for one of several possible reasons. Although these communication failures do not necessarily mean the queried e-mail is invalid, the failures were relatively consistent across surveys and are thus counted as errors.

hubs may disproportionately affect the number of e-mail addresses that are lost. The domains with the most associated CAEs in MEDLINE were commercial (in order of descending prevalence): hotmail.com, yahoo.com and aol.com.

We further compared the number of valid and invalid e-mail addresses by their top-level domain—for example, 'edu' is the top-level domain in 'jsmith@genetics.ou.edu'—while excluding e-mails in 'accept all' domains and domains with which communication could not be established. We found that the overall probability that a CAE is valid varies significantly by country and/or organization. Restricting the analysis to countries and groups with at least 2,000 domain names (Table 2), we found that CAEs from Poland are the most stable (21% invalid) and CAEs from the Netherlands are the least stable (53% invalid). Expanding the analysis to include countries and groups with at least 100 domain names, the least stable is Singapore (89% invalid) and the most stable is Lithuania (9% invalid). The reason behind these country-level differences is probably because of socio-cultural policies on how corresponding authors are appointed. For example, in some universities and/or countries, it may be customary for the senior author or lab mentor to assume the responsibility of communication for a body of work. In this case, the CAE will probably

remain available for a longer period than it would in a system in which a student—who may be closest to the body of research—is appointed as CA. In the latter case, students normally graduate and leave their institutions within much shorter time frames than their mentors.

We then sought to determine whether stability is correlated with the number of sub-domains present in an e-mail address. CAEs were classified as invalid, unable to determine due to communication failure, or OK (valid). As before, only e-mails that did not fall into the 'accept all' category were included. The probability of an e-mail address becoming invalid after publication correlates with the number of sub-domains present (e-mails with no top-level domain are all invalid; Table 3). This makes sense because higher-level e-mail servers (for example, jsmith@ou.edu) tend to be supported by more personnel and funding, whereas departmental (for example, jsmith@biochemistry.ou.edu) or sub-departmental (for example, jsmith@smithlab.biochemistry.ou.edu) servers tend to be dependent on fewer individuals and less funding. Interestingly, the probability that the CAE is invalid due to syntax errors or individual user account failure does not appear to correlate with the number of sub-domains present.

Spelling errors—made by the author, the journal or personnel at the National Library of Medicine—invalidated some e-mail addresses. For example, we found

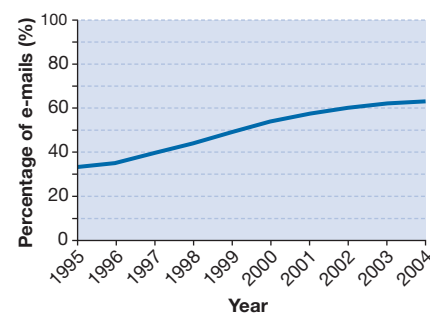


Fig 3 | Growth in the percentage of e-mail addresses in a domain that accepts any e-mail address as valid according to the year that the domain first appeared in MEDLINE.

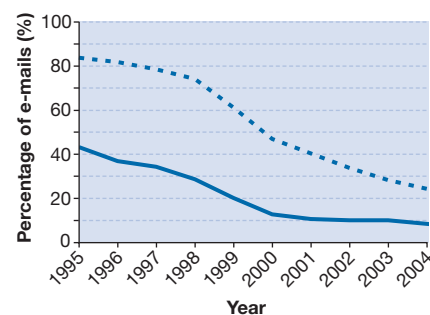


Fig 4 | Percentage of author e-mail addresses returning an error code. Those published in MEDLINE that were not valid when queried in October/November 2004 are shown as a dotted line. The year refers to the date of their first appearance in MEDLINE. The solid line represents addresses not valid due to a bad domain.

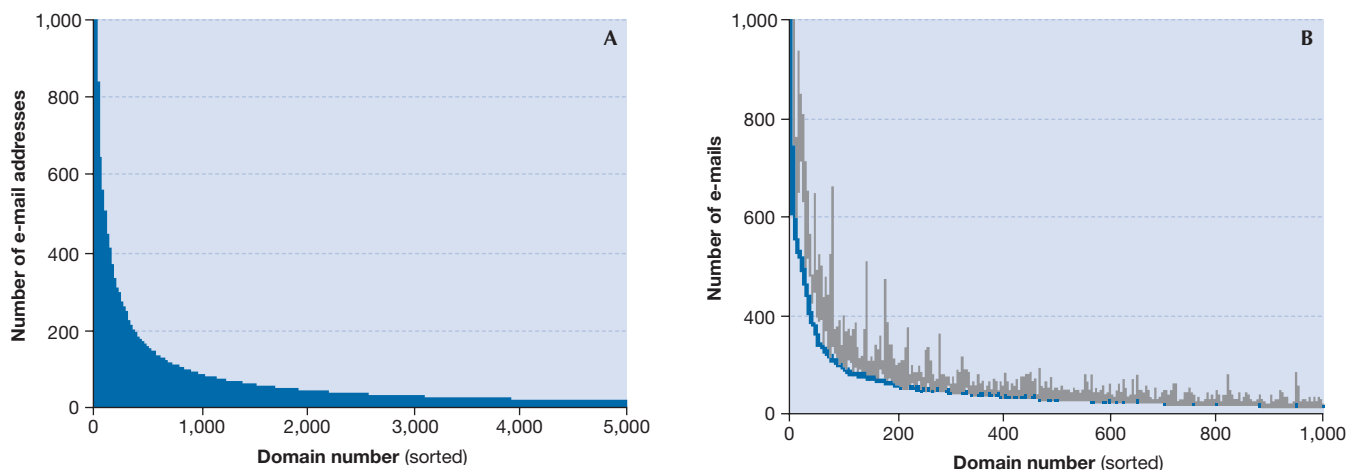


Fig 5 | Number of corresponding author e-mails associated with domain names. **(A)** The number of corresponding author e-mails (CAEs) associated with the 5,000 domains published most frequently, and **(B)** distribution in the number of CAEs attached to domains that have become invalid (blue line) and the number of times an e-mail address in that domain was published (grey line). The y-axis in **(B)** is limited to intervals of 200 so the trend is visible and only the first 1,000 records are shown, sorted in descending order.

25 spelling variations for the most frequently published e-mail domain, u.washington.edu, which represented 1.5% of the e-mail addresses with this domain. Although this is a relatively small fraction of the total e-mails published, it is apparent that PubMed/MEDLINE would benefit from some form of quality control. Simple checks, such as the kind performed here for domains and e-mail addresses, would catch approximately 1% of the errors. Other errors may not be obvious until someone attempts to e-mail the published address.

E-mail correspondence is becoming increasingly common within biomedical fields, but CAEs expire rapidly, with 45% of published e-mail addresses invalid at the time of this study. Almost one-third of this decay was due to the loss of the e-mail server (domain) hosting the account. This rate of decay could serve as a guideline for the number of participants necessary to conduct a survey, should researchers choose e-mail as a primary means of contacting participants. For example, if researchers anticipate that 1,000 responses are necessary for statistical significance, they should begin with a minimum of 2,000 e-mail addresses for a 5-year follow-up and 4,000 for a 10-year follow-up. Our study also revealed that CAE stability correlates with top-level domain and that CAEs originating from some countries or organizations may be

more stable than others. One thing that cannot be accounted for within this study is the number of e-mail accounts that are still active on their domain servers but are no longer accessed by their users.

One in four e-mail addresses becoming invalid within one year of publication is an alarming rate of decay as it has an impact on the ability of scientists to communicate and exchange material. We therefore have several simple suggestions for policy changes that could help to enhance the continuity of e-mail contact. First, and perhaps simplest, journals and/or curators of citation databases should check e-mail addresses for proper syntax and valid domains, although this step, unfortunately, would not prevent more than a small fraction of e-mail deactivation. Second, also simple yet liable to have far greater impact, journals should publish and citation database curators should index all co-author e-mails, which would serve as backup contact information if the CAE becomes invalid. Presumably, co-authors should be able to answer some questions about the published research and are among the people most likely to know the whereabouts of the CA. In a time when the number of authors per paper is growing steadily, this step not only seems more likely to succeed in preserving contact but also has the advantage of linking more author names, which are not necessarily unique, to e-mail addresses, which must be unique, to more easily disambiguate author names.

Table 2 | Stability of corresponding author e-mails by top-level domain, sorted by the percentage that are invalid

TLD	Country/Type	Valid	Invalid (%)
.nl	Netherlands	2,485	2,802 (53)
.ca	Canada	3,224	3,008 (48)
.jp	Japan	6,498	5,880 (48)
.org	Organization	1,878	1,562 (45)
.ch	Switzerland	1,784	1,482 (45)
.de	Germany	9,722	7,873 (45)
.no	Norway	1,395	1,123 (45)
.gov	Government	1,796	1,423 (44)
.fr	France	4,294	3,148 (42)
.br	Brazil	2,198	1,555 (41)
.com	Commercial	9,049	6,247 (41)
.uk	UK	13,970	9,596 (41)
.au	Australia	3,409	2,229 (40)
.se	Sweden	2,525	1,562 (38)
.es	Spain	1,956	1,201 (38)
.net	Network	3,961	2,267 (36)
.edu	Educational	38,500	19,506 (34)
.dk	Denmark	1,756	869 (33)
.it	Italy	7,114	3,138 (31)
.fi	Finland	2,359	792 (25)
.at	Austria	1,985	617 (24)
.pl	Poland	1,989	533 (21)

TLD, top-level domain.

Third, corresponding authors should provide two e-mail addresses, one of which is institution-independent, to provide an alternative means of contact if they move

Table 3 | Corresponding author e-mail stability correlates with the number of sub-domains within the address

E-mail account status	Number of sub-domains					
	0*	1	2	3	4	5+
Total	505	87,132	103,080	35,884	5,271	187
OK	–	61,719	61,167	16,280	1,710	36
Invalid account (but domain OK)	505	20,029	24,049	8,134	1,038	46
Domain failure	–	5,384	17,864	11,470	2,523	105
Communication failure	–	7,134	9,132	5,957	747	14
Invalid (but domain OK) (%)	100	23	23	23	20	25
Bad due to domain failure (%)	0	6	17	32	48	56
Communication problems (%)	0	8	9	17	14	7

*E-mails without a top-level domain are always invalid.

One in four e-mail addresses becoming invalid... is alarming as it impacts on the ability of scientists to communicate and exchange material

or if their institutional or departmental e-mail server becomes inactive. Institution-independent e-mail hosting sites such as Microsoft's Hotmail or Yahoo's services provide a means of contact continuity when authors change jobs, and they are already widely used. But these could become a point of critical failure as the number of users grows; for example, if Microsoft began charging for its Hotmail e-mail accounts, many users might switch to another service. Fourth, but requiring a much greater level of diligence, journals and reviewers should be aware that some domains are less stable than others, and on seeing statistically unstable top-level domain names or deep sub-domain structures, could request the authors to provide additional contact details. Finally, but more intrusive on personal liberties, departments and/or institutions should dissuade employees from creating their own e-mail servers and ask that they rely instead on e-mail servers with more persistent means of support. Given that e-mail—a highly efficient and increasingly important means of

communication—has gained such prominence in daily research work, anything that could be done to ease contact between scientists surely deserves attention.

ACKNOWLEDGEMENTS

This work was funded in part by the National Science Foundation, grant EPS-0447262.

Supplementary information is available at http://faculty-staff.ou.edu/W/Jonathan.D.Wren-1/Papers/Email_study_All_runs.xls.

REFERENCES

- Cross TB (1983) The grapefruit diet. *J Microgr* **16**: 14–18
- Dellavalle RP, Hester EJ, Heilig LF, Drake AL, Kuntzman JW, Graber M, Schilling LM (2003) Going, going, gone: lost Internet references. *Science* **302**: 787–788
- Dodds PS, Muhamad R, Watts DJ (2003) An experimental study of search in global social networks. *Science* **301**: 827–829
- Hardy IR (1996) *The Evolution of ARPANET Email*. Thesis, Department of History, University of California, Berkeley, CA, USA
- Knauper B, Rabiau M, Cohen O, Patriciu N (2004) Compensatory health beliefs: scale development and psychometric properties. *Psychol Health* **19**: 607–624
- Lawrence S, Pennock DM, Flake GW, Krovetz R, Coetzee FM, Glover E, Nielsen FA, Kruger A, Giles CL (2001) Persistence of web references in scientific research. *IEEE Comput* **34**: 26–31
- Newman ME (2004) Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci USA* **101** (Suppl 1): 5200–5205
- Nguyen DT, Murphy J (2001) *Australian Organisations' Email Customer Service*. Proceedings of the 4th Western Australian

Workshop on Information Systems Research. Perth, WA, Australia: University of Western Australia

Oppermann M (1995) E-mail surveys—potentials and pitfalls. *Marketing Res* **7**: 28–33

Pallen M (1995) Electronic mail. *Br Med J* **311**: 1487–1490

Singarella T, Baxter J, Sandefur RR, Emery CC (1993) The effects of electronic mail on communication in two health sciences institutions. *J Med Syst* **17**: 69–86

Spinellis D (2003) The decay and failures of web references. *Commun ACM* **46**: 71–77

Thach L (1995) Using electronic mail to conduct survey research. *Educ Technol* **35**: 27–31

Torvik VI, Weeber M, Swanson DR, Smalheiser NR (2003) A probabilistic similarity metric for Medline records: a model for author name disambiguation. *AMIA Annu Symp Proc* **1033**

Treadwell JR, Soetikno RM, Lenert LA (1999) Feasibility of quality-of-life research on the Internet: a follow-up study. *Qual Life Res* **8**: 743–747

Wren JD (2004) 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics* **20**: 668–672

Wren JD (2005) Open access and openly accessible: a study of scientific publications shared via the internet. *Br Med J* **330**: 1128

Jonathan D. Wren, Joe E. Grissom and Tyrrell Conway are at the Advanced Center for Genome Technology in the Department of Botany and Microbiology at the University of Oklahoma, Norman, OK, USA.

E-mail: jonathan.wren@ou.edu (jdwren@gmail.com); jgrissom@ou.com (jegrissom@gmail.com); tconway@ou.edu

doi:10.1038/sj.embor.7400631