Analyzing Implicit Group Messaging: a novel messaging paradigm for group-oriented content distribution

Neil Cowzer, Paddy Nixon
School of Computer Science and Infomatics
UCD Casl
Belfield Office Park, Dublin 4
Email: {neil.cowzer, paddy.nixon}@ucd.ie

***Abstract*—Publish-subscribe systems are well suited loosely decoupled nature of the web, resulting in the messaging paradigm gaining widespread adoption and being the subject of much research. Such research has focused primarily on architectures and filtering algorithms with little evidence of performance analysis or characterization of user behavior in these widely deployed messaging paradigms. In this paper we discuss and examine implicit group messaging; an application-layer many-to-many messaging paradigm for delivering messages from publishers to specified groups of consumers. Such consumer groups are not addressed by explicit names, instead they are reached by describing the shared attributes or interests of consumers, forming easily defined implicit groups.**

Based on a 4 week experiment we analyze the characteristics of implicit groups and their usage. We find implicit group messaging workload to be similar to RSS in terms of group membership and update patterns; groups are typically small with few large examples and update rates vary from infrequent to more limited intervals.

I. Introduction

Publish-subscribe, or *pub-sub*, systems come in a number of varieties [1] and are gaining widespread acceptance [2]–[8]. Previous research in this area has focused primarily on aspects such as: system architectures, event notification and filtering algorithms. Limited research has tackled the pertinent issue of characterizing publisher and subscriber behavior in pub-sub and its variants.

This paper examines this behavior with regard to implicit group messaging (IGM); a novel messaging paradigm based on the formation of implicit groups by the intersecting interests of users. Perhaps most easily visualized as a form of pub-sub, IGM requires users subscribe to content by signaling their interests, e.g. a series of tags. Publishers then *select* the recipients for content by describing the shared interests of the intended audience through a primitive language. Key to IGM is assumption that the publisher knows (and can describe) the intended audience for their content, with recipients selected by delimiting their shared attributes or interests. This represents a contrast to "traditional" pub-sub where publishers do not select recipients, instead recipients select the publishers or types of content they wish to consume. Defining groups in this way results in the addition of a level of semantics otherwise missing in alternative messaging paradigms, with message subject easily applied to and extracted from implicit group definitions. IGM is a flexible message paradigm and can be applied to a variety of fields. In this paper, we limit our discussion of IGM to that of web micro-news distribution, similar to that of RSS.

Until now there has been no real-world example of an IGM system, only simulations. It follows that there has been no evaluation of user and group behavior on implicit group platforms. Educated estimates of group size and distribution were extracted from other pub-sub systems and utilized for previous simulations. We verify many of these assumptions.

Our study provides several insights into IGM. We find IGM group size follows a heavy tailed, Zipf distribution similar to that of RSS [9]. The majority (80%) of groups formed are small with very few (4%) large groups exists.

Second, IGM groups exhibit extreme variation in publishing rate, with some groups receiving many updates per day, others receiving a single update throughout the duration of the experiment.

The formation of (implicit) groups at publishing time enables IGM service transient interests such as product announcements or significant events. We demonstrate this by noting the impact of 3 significant events were observed in the RSS feeds we monitored, namely: the launch of the Google Nexus, Consumer Electronics Show and the launch of Apple's iPad.

This paper makes the following contributions. We demonstrate an application of implicit group messaging in the form of web micro-news distribution. We characterize the behavior of groups and users, contrasting their performance to that of generic RSS feeds. Finally, we discuss the impact of our results and propose future extensions to implicit group messaging platforms.

II. Background and Related Work

Publish-subscribe systems are a well researched topic. In this section we provide a brief introduction to the field and discuss related material.

978-1-4244-8865-0/10/$26.00 ©2010 IEEE

Authorized licensed use limited to: University of Canberra. Downloaded on June 17,2020 at 23:37:47 UTC from IEEE Xplore. Restrictions apply.

Pub-sub

The pub-sub interaction paradigm enables subscribers to express their interest in an event or series of events, with a view to being notified upon the occurrence of any subsequent event(s) that match their registered interest. The general flow of pub-sub systems requires subscribers specify their interest by subscribing to a *feed*, also known as *topic*, *channel*, *subject* or *group* (herein we use these interchangeably). A system monitors content or *events* produced by publishers, matching against users subscriptions and distributing events accordingly.

In general, pub-sub systems are classed in terms of their expressiveness, with the majority of systems falling into one of two classes: topic-based and content-based. In topic-based pub-sub publishers events are labeled with subjects and these labels are then compared to subscriptions. Our comparison messaging paradigm, RSS [10], is the best-known example of this topic-based pub-sub. Content-based approaches differ, relying on attributes of the content to match subscriptions made by subscribers.

Implicit Group Messaging

Implicit group messaging, or IGM, is a novel messaging paradigm which serves content from publishers to specified implicit groups of users [11]. Underpinning this concept is the notion of an implicit group: a set of consumers that have some inherent features in common. As a consequence, implicit groups are addressed by defining the characteristics of its members, rather than by explicit names.

While distinct, IGM is perhaps easiest conceptualized as a form of pub/sub system, in which users are only required to signify their interest in set of topics and publishers select the appropriate recipients by describing their interests in the form of *target-expressions*. This can be seen as a reversal of roles in relation to the typical pub/sub systems where, in practice, users are the selectors of content. Furthermore, implicit groups are highly dynamic; implicit groups are only defined at the time of publication. It follows that the actual membership of an implicit group may vary significantly from message to message as participants join and leave the system or as their attributes change over time. Figure 1 illustrates message dissemination in IGM, only consumers that satisfy the combinatory interests receive the message.

IGM does not does not specify any particular modeling language to describe consumers or implicit groups, choosing to allow a suitable language may be chosen for differing domains. In [11], Cutting utilizes an limited Boolean based language, which we make use of for this experiment:

$$tag := [a - zA - Z]+ \qquad (1)$$

$$expression := item(``\mid"item \mid ``\&"item)* \qquad (2)$$

$$item := tag \mid ``("expression``)" \qquad (3)$$

The described language utilizes limited logic operations; disjunction | and conjunction &, to formulate target expressions. For example, a user interested in Football World Cup would generate the following expression; (“$Football$”&“$WorldCup$”).

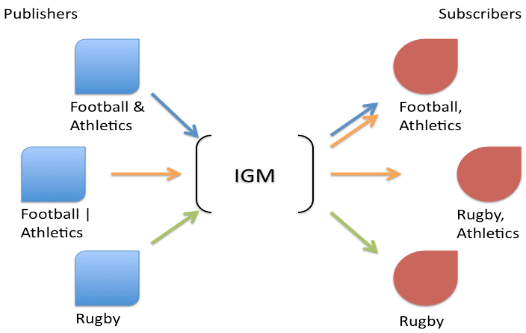

Fig. 1. Message flow in IGM, only consumers with that satisfy a messages target-expression receive the messages from publishers.

A. Related Work

The majority pub-sub research has focused primarily on the design and implementation of scalable efficient event dissemination platforms [2], [3], [11]–[14]. Scaling and efficiency issues also arise in RSS, particularly with regards to subscriptions. Accordingly a number of proposals have been put forth; FeedTree [7], Corona [15], RSSCloud (part of RSS specification [10]) and PubSubHubbub [16]. All these approaches try to minimize polling and/or update latency through the cooperation of participating nodes. RSSCloud and PubSubHubbub also add the ability for “true” push notifications encouraging timely updates. Fethr [8] operates on similar principles, however Fethr is specialized for micro-publishing services such as Twitter precluding it from further discussion.

IGM is currently in the nascent stages and has not garnered the widespread adoption of RSS. With no pre-existing IGM systems no real-world measurement study exists, however authors expected performance characteristics similar to that of RSS. The authoritative paper in this field is Liu et. al [9], the first measurement study is of RSS or any Internet-scale pub-sub system. Liu found RSS loads to represented by a heavy-tailed Zipf distribution, similar to that of web objects. This paper represents the first measurement study of a real-world IGM system.

III. Experiment Setup

Until now there has been no real-world example of an IGM system, only simulations. It follows that there has been no evaluation of user and group behavior on implicit group platforms.

We conducted an experiment contrasting the performance of IGM against RSS for web micro-news dissemination. The experiment was conducted over a 4 week period, with 21[1] participants. Upon registration, users selected 10 out of 20 popular RSS feeds and provide a list of freely selected tags delimiting their interests in these feeds. Over the duration of

[1] 31 participants initially registered to utilize the system, however due to user inactivity their results were omitted from our results

Authorized licensed use limited to: University of Canberra. Downloaded on June 17,2020 at 23:37:47 UTC from IEEE Xplore. Restrictions apply.

the experiment, 4137 articles were collected and disseminated through both messaging paradigms. Each week a participant received two feeds reviewing the highest ranked articles collected during the past week: a feed generated through 10 standard RSS subscriptions and an alternative utilizing IGM and the tagged interest of users. News items were ranked according to their occurrences on social networks such as Twitter[2] and Facebook[3] as calculated by Bit.ly[4].

Subscriptions were limited to 20 popular technology feeds to ensure cross-over of user tags (thereby enabling implicit group formation). The relatively high number of subscriptions, 10, ensured the participants were unlikely to remember which feeds they specified [17], an effect exacerbated by the week spacing between registration and active participation. To remove user bias towards specific publishers, articles were initially presented with limited information: title and description, identifying information such as url and publisher were omitted however they were enabled at a later point to allow access to full articles.

Constructing implicit groups

While obtaining the sources for the RSS component are widely available, there exists no source of content with target-expressions specified according to the language defined in section II. As a result, suitable target-expressions must be retro-fitted to articles and sources. By utilizing only RSS feeds including tag descriptions of articles, target-expressions could be constructed based on the publisher provided tags. To generate target-expressions we made use of two simple assumptions:

- authors infer weighting of tag relevance through order, i.e. the first tag is most relevant, second is second most relevant …
- tags are independent, i.e. a tag does not rely on a preceding tag for context.

Recursively, we removed the least important tag (appending to list e) and evaluated our target expression s against users tag lists. Upon success, we allowed an arbitrary number b of iterations to enable the creation of expressions containing the disjunction operator. This process can be seen more clearly in the pseudo code III.1.

Algorithm III.1: EXPRESSION(s, e, b)

```
for each $u \in \mathcal{U}$
  do if SELECTS$(s, u)$
  then $exp \leftarrow s$
if LEN$(s) > 1$ and LEN$(exp) < b$
  then for $i \leftarrow 0$ to LEN$(e)$
    do { $s_1 \leftarrow s[:-1] + e[i]$
       { $e_1 \leftarrow e[:i] + e[i+1:] + s[-1]$
       { $exp \leftarrow$ EXPRESSION$(s_1, e_1, b)$
return $(exp)$
```

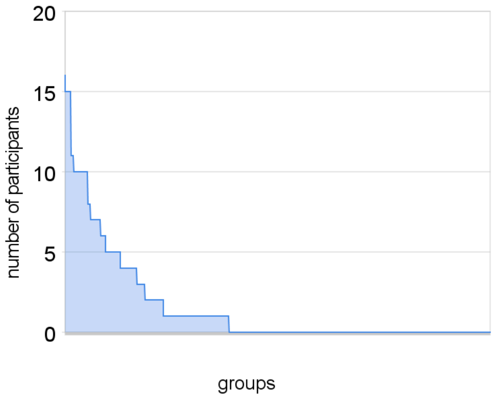

Fig. 2. IGM groups ranked by the number of participants follows a Zipf distribution.

An example output of the algorithm may be:

$$A\&B, A\&C\&D, A\&E \tag{4}$$

where A,B,C,D and E are independent tags. This can be written more compactly as:

$$A\&(B \mid (C\&D) \mid E) \tag{5}$$

and therefore adheres to our defined language.

Initially we experimented with relatively small values for b, however content propagation issues arose in our system as articles were only distributed to a few highly tagged individuals. Utilizing larger values for b, thereby allowing more generic implicit groups, resolved this problem.

IV. Survey Results

We report on two aspects of the IGM system using the experimental setup described above. First, analyzing the characteristics of implicit groups, such as their popularity distribution and update rate. Second, we investigate user behavior in implicit groups; their click-through rates and patterns in tagging. We ground these results with comparisons to the RSS subscription method employed.

Group Behaviour

Figure 2 shows the popularity of implicit groups ranked by the number of users contained. A log-log trace of this data confirms group-size follows a Zipf distribution, incurring an α of 1.88. A long tail is evident with many niche groups receiving 3 or less participants. Similar to other web phenomenon, there exists the presence of relatively few large groups, which for the purpose of this experiment we class as 10 or more members, while there are many small or niche (less than 5 members) groups, representing 4% and 80% of the total groups, respectively.

In figure 3, update rate is shown to follow a similar Zipf distribution with groups falling into two distinct update patterns; either they are a) updated many times or b) they are rarely updated. This is borne out in the statistics; 73% of groups receive less than 10 updates in the 4 week period

[2] http://www.twitter.com/

[3] http://www.facebook.com/

[4] http://www.bit.ly/

Authorized licensed use limited to: University of Canberra. Downloaded on June 17,2020 at 23:37:47 UTC from IEEE Xplore. Restrictions apply.

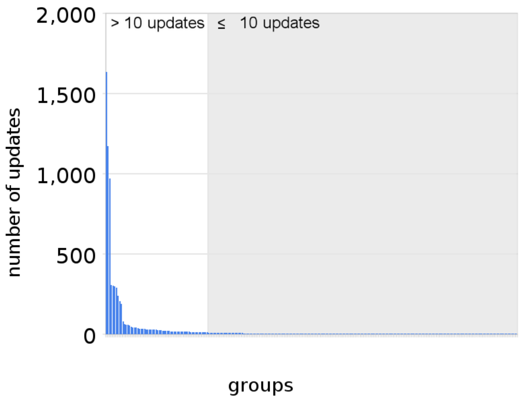

Fig. 3. Total updates per (implicit) group over the 4 week period.

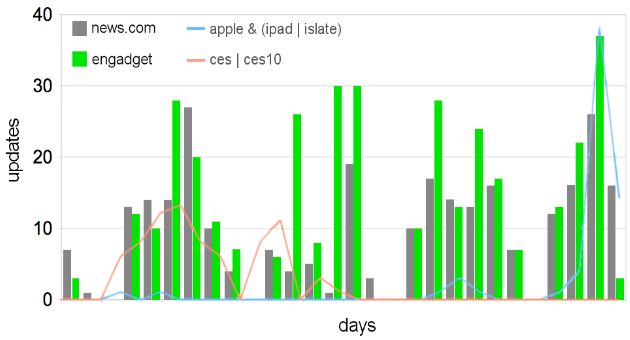

Fig. 4. Events observed in IGM and RSS; igm groups are good indicators of "trending" topics or events.

whereas only 4% of groups receive more than 100 updates. This result suggests that the majority of groups are not too resource intensive, however a minority of these groups are likely to create issues due to their scale, albeit groups on a larger scale than our experiment.

However, it may not be so simple, the number of updates per group varies significantly over time. Events such as Internet memes, holidays and product announcements are responsible for the creation of new groups and variations in their update rate. For example, the Consumer Electronics Show (CES) took place in January 1st to 7th and Apple's iPad launched on January 27th. In figure 4, we plot the trace the update rate for the 2 implicit groups related to these events against the two of the most popular RSS feeds. The events are clearly shown to correspond to significant increase in two RSS feeds presented, adding a level of semantics previously unavailable to web micro-news distribution.

User Behaviour

Implicit groups are formed by the overlapping interests, in our case tags, of users. As such tagging behavior significantly impacts our system. Total number of tags per user ranged between 2 and 17 with a mean and mode of 9.5 and 8 respectively. On average a users is a member of 22 of these groups; maximum and minimum memberships of 68 and 4 are recorded. Surprisingly these do not correspond to the most (17) and least (2) tagged individual. Instead users who utilize more common or popular tags, are likely to have more in common with users. Simply put, it matters more what you say rather than how much you say. Given the technology theme of the potential RSS feeds, there is no surprise to see, companies and product names such as Apple, Google, Microsoft and Facebook being the most frequent entities accounting for 7 of the top 10 most frequent tags.

The specificity of user tags plays a pivotal role in update rate of content they receive. In general, generic tags such as brand names, e.g. Apple, Google and Microsoft, encountered higher update rates than more specific tags such as products names, e.g. iPod Touch, Android and Windows 7. Users displayed tendency to utilize generic tags, demonstrated by the company names dominating the list of most frequently used tags.

IGM outperformed, our comparison messaging paradigm, RSS in terms of the articles read, with 11% increase in total number of articles read. Significantly, these articles cater for a wider variety of sources and interests — an indicator that IGM is servicing some niche groups — with 150 distinct articles, an improvement over 96 presented by RSS.

IGM performed similar to previous estimates for group size and distribution, with heavy-tailed Zipf-like distributions similar to that of RSS experienced [9]. Furthermore, IGM was shown improve both content distribution and user interaction. While our results are limited by the small user group size, we expect them to generalize well to larger audiences and thus represent a good basis for future work.

V. Conclusion and Future Work

This paper presents a measurement study of IGM, a messaging paradigm for serendipitous content discovery. Until this experiment, no real-world evaluation of IGM had existed, as such it provides key insights into how an IGM system is utilized in practice and highlighting issues for the design of future IGM systems.

A significant focus of our study was to analyze the group formation and behavior in IGM. This study shows that IGM systems operates in a similar manner to other pub-sub systems, encountering Zipf-like distributions. Implicit group update rate is shown to vary significantly posing a challenge for systems design. Furthermore, although our experiment setup did not allow it, we can assume that the increases in update rates are likely to correspond to an increase in group membership (similar to flash-crowds around a popular or significant event) thereby exacerbating the problem. Group membership is also shown to be distributed unevenly verifying previous assumptions. The majority of groups (80%) contain 3 or less participants, and the remaining 20% of groups having between 4 and 10 members. While these figures for group size do not cause alarm, on a larger scale such a distribution will present major difficulties for service providers.

A fundamental goal of IGM is to provide so-called serendipitous content discovery; by this we mean the discovery of

Authorized licensed use limited to: University of Canberra. Downloaded on June 17,2020 at 23:37:47 UTC from IEEE Xplore. Restrictions apply.

relevant content that one would of otherwise have missed. Although both methods utilized the same set of articles, they present very different subsets of articles to users. Of the 150 distinct articles presented to users by IGM, 116 (77%) of these were unique to IGM. Meanwhile just over 64% (61 of 96) articles presented by RSS were unique, ensuring that both methods satisfying our description of serendipity. However we argue that IGM does so in a more meaningful way. IGM ensures content is more closely aligned to users interests. Furthermore IGM avoids the *long-tail* problem. In RSS, 96 distinct articles were encountered, then weighted according to our external ranking value and presented to the users. Unsurprisingly, articles and topics with an high external ranking value tended to dominate the lists presented to users, regardless of how relevant the articles topic would be to a user's interest. In effect, mob rule ruled, with the top articles being the ones of relevance to the largest amount of people (and not the user in question). Little personalization exists - niche interests are not catered for. IGM differs slightly, user interest is first and foremost, enabling the formation of niche groups and therefore an increase in the number of tastes catered for. This can been seen in a simple comparison of the total number of distinct articles presented, 150 (IGM) to 96 (RSS).

Despite the marked improvement displayed by IGM, there is evidence that otherwise relevant content "miss" their intended audience due to inconsistencies in publisher/user tagging behavior. It is interesting to note, our retro-fitted target-expressions represent a best-case scenario, formed with universal knowledge of user and publisher tagging behavior leading to the creation complex target-expressions reaching the maximum available audience. Furthermore, while defining groups with a small number of interests represents an approachable model for publishers, definitions can quickly become complex as the publishers attempt to produced more fine-grained groups. In this regard, we plan to explore social evolving descriptors utilizing peer knowledge to amend event target-expressions. For example: imagine when a user reads articles tagged "apple" and "ipod", they are generally also tagged "iphone". The discussed user reads an article tagged by the author with "apple" and "iphone" only, perhaps they even signifies their approval of the article. There is a high chance that the *missing* "ipod" tag is relevant to this article. We therefore recommend this tag be appended to the article description and propagate the updated expression to the relevant parties. In this way users can influence the articles circulation and the burden on publishers to described content fully and correctly is alleviated.

Requiring users to tag their interests for a small field such as technology related news feeds is a trivial task, placing little burden on the user. Doing so for larger fields may represent an excessive burden on the user. In future work, we hope to resolve this problem. Inherent in social evolving descriptors is the learning of user interests by way of observed tag relations. In effect, we are maintaining user profiles, which can be utilized to modify and refine user tag subscriptions: new interests or tags may be added to users profiles or due falling interest, others may be omitted.

In summary, this paper represents the first study of real-world usage an IGM system. The performance characteristics of IGM are outline, issues highlighted and potential solutions offered. We hope this study will aid designers to understand, design and evaluate future IGM systems.

VI. Acknowledgments

This work is funded as part of the IRCSET Embark initiative[5].

The authors would like to thank Clarity[6] members Peter Briggs, Maurice Coyle and Owen Phelan whose input to the experiment setup and evaluation was immeasurable.

References

[1] P. Eugster, P. Felber, R. Guerraoui, and A. Kermarrec, "The many faces of publish/subscribe," *ACM Computing Surveys (CSUR)*, vol. 35, no. 2, p. 131, 2003.

[2] N. Carriero and D. Gelernter, "Linda in context," *Communications of the ACM*, vol. 32, no. 4, p. 458, 1989.

[3] M. Castro, P. Druschel, A. Kermarrec, and A. Rowstron, "SCRIBE: A large-scale and decentralized application-level multicast infrastructure," *IEEE Journal on Selected Areas in communications*, vol. 20, no. 8, pp. 1489–1499, 2002.

[4] L. Fiege, F. Gartner, O. Kasten, and A. Zeidler, "Supporting mobility in content-based publish/subscribe middleware," *Lecture Notes in Computer Science*, pp. 103–122, 2003.

[5] P. Eugster, "Type-based publish/subscribe: Concepts and experiences," *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 29, no. 1, p. 6, 2007.

[6] D. Cutting, B. Landfeldt, and A. Quigley, "Implicit group messaging over peer-to-peer networks," in *P2P '06: Proceedings of the Sixth IEEE International Conference on Peer-to-Peer Computing*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 125–132.

[7] D. Sandler, A. Mislove, A. Post, and P. Druschel, "Feedtree: Sharing web micronews with peer-to-peer event notification," *Lecture notes in computer science*, vol. 3640, p. 141, 2005.

[8] D. Sandler and D. Wallach, "Birds of a fethr: Open, decentralized micropublishing," in *8th International Workshop on Peer-to-Peer systems (IPTSP'09)*, Boston, MA, April 2009.

[9] H. Liu, V. Ramasubramanian, and E. Sirer, "Client behavior and feed characteristics of rss, a publish-subscribe system for web micronews," in *Proc. of ACM Internet Measurement Conference*, 2005.

[10] "Rss 2.0 specification," July 2003, http://cyber.law.harvard.edu/rss/rss.html.

[11] D. Cutting, A. Quigley, and B. Landfeldt, "Special Interest Messaging: A Comparison of IGM Approaches," *The Computer Journal*, 2007.

[12] A. Carzaniga, D. Rosenblum, and A. Wolf, "Design and evaluation of a wide-area event notification service," *ACM Transactions on Computer Systems (TOCS)*, vol. 19, no. 3, pp. 332–383, 2001.

[13] IBM, "Tspaces - computer science research at almaden," http://www.almaden.ibm.com/cs/TSpaces/.

[14] E. Freeman, K. Arnold, and S. Hupfer, *JavaSpaces principles, patterns, and practice*. Addison-Wesley Longman Ltd. Essex, UK, UK, 1999.

[15] V. Ramasubramanian, R. Peterson, and E. Sirer, "Corona: A high performance publish-subscribe system for the world wide web," *Proceedings of Networked System Design and Implementation (NSDI)*, 2006.

[16] B. Fitzpatrick, B. Slatkin, and M. Atkins, "Pubsubhubbub core 0.2 – working draft," September 2009, http://pubsubhubbub.googlecode.com/svn/trunk/pubsubhubbub-core-0.2.html.

[17] G. Miller *et al.*, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological review*, vol. 63, no. 2, pp. 81–97, 1956.

[5] http://www.ircset.ie/

[6] http://www.clarity-centre.org/

Authorized licensed use limited to: University of Canberra. Downloaded on June 17,2020 at 23:37:47 UTC from IEEE Xplore. Restrictions apply.