A Case for Information-Bound Referencing

Ashok Anand, Aditya Akella
Univ. of Wisconsin-Madison
Madison, WI
{ashok,akella}@ cs.wisc.edu

Vyas Sekar
Intel Labs
Berkeley, CA
vyas.sekar@intel.com

Srinivasan Seshan
Carnegie Mellon University
Pittsburgh, PA
srini@cmu.edu

ABSTRACT

Links and content references form the foundation of the way that users interact today. Unfortunately, the links used today (URLs) are fragile since they tightly specify a protocol, host, and filename. Some past efforts have decoupled this binding to a certain degree; e.g., creating links that bind to byte-level data. We argue that these systems do not go far enough. Our key observation is that users really care about the intent of the referenced link and are relatively agnostic to the byte-level representation. Based on this observation, we argue that references should be bound to the underlying information associated with the referenced content. We call such references Information-Bound References (IBR). In this paper, we focus on the challenges of creating IBRs for multimedia data, since these form a dominant fraction of Internet traffic today. We explore the trade-offs of various alternatives for generating and using IBRs. We identify that it is possible to adapt multimedia fingerprinting algorithms in the literature to generate IBRs.

Categories and Subject Descriptors

C.2.1 [**Computer Communication Networks**]: Network Architecture and Design

General Terms

Design

1. INTRODUCTION

Links and content references form the foundation of the way users interact with the Internet today. Unfortunately, as past work has noted [17], the links that are in wide use today (i.e., URLs) are fragile since they tightly specify a specific protocol, host and file name. Past efforts have made the observation that users can be agnostic to the protocols, hosts or file names involved. Some of these efforts [17] remove the tight binding of a link to a host and others [29, 28] go further and replace the link to protocol, host, and file name with a binding to the byte-level data. While these solutions take an important step toward making links more robust allowing users to fetch content from anywhere using any transfer protocol, we argue that these systems do not go far enough. The key observation that allows us to push the design space further is that human users are the actual final consumers of content. Users are less concerned with the byte level representation of data as long as the content retrieved matches the *intent* of the referenced link.

Based on this observation, we believe that references should *not* be bound to protocols, hosts, file names or underlying data. Instead they should be bound to the underlying *information* associated with the referenced data. Note that by *information*, we are referring to a "perceptual" entity that is invariant across different presentation formats, encodings, and resolutions. Whereas, by *data*, we refer to information coupled with a specific presentation instance. We call such references *Information-Bound References (IBR)*. In this paper, we focus on the challenges of creating IBRs for multimedia data since transfers for such data are the largest contributor to Internet traffic.

IBRs can be a key enabler for both publishers and consumers of multimedia content. In order to accommodate heterogeneity in user device capabilities today, publishers of multimedia content need to ensure that they support multiple possible formats and resolutions—a daunting task especially for user-generated content. Furthermore, multimedia items are often already available from several locations and in multiple formats and resolutions. However, in the restrictive model of URL based references, consumers have no way today to leverage this broader availability of content. Consumers are often stuck with formats that are not compatible with their device. Publishers cannot automatically direct consumers to alternate third-party sources who may have the appropriate format. Even if content publishers were to integrate with such third parties (e.g., using P2P CDNs [18]), there is no way to verify the integrity of content. Using IBRs, content publishers can *easily* reach a wider target audience as well as provide high availability and easy integration with

HotNets'10, October 20–21, 2010, Monterey, CA, USA

third party services. Consumers can readily explore different variants of content, and choose the most appropriate fit for their operating constraints.

Two sets of mechanisms are required for enabling IBRs: (i) those for generating IBRs and (ii) those for resolving IBRs to find a matching piece of content on some host. IBR resolution should allow consumers to *negotiate content properties* by specifying IBRs along with their constraints (e.g., bandwidth or rendering capabilities). It should also allow content publishers to *register* their ability to provide content corresponding to an IBR in some specific format.

The IBRs generated for any piece of information must meet several key requirements. They must be encoding-, resolution- and location-independent. To ensure contention-freeness and stability of IBRs, they must be algorithmically derived from the content based on the information it carries (as opposed to, e.g., human-input labels such as "*Charlie bit my finger*"). IBRs must be unique in that IBRs for two different pieces of content should never match. Finally, it should not be possible to create content that matches a specific IBR.

Note that these requirements partially match the requirements of the cryptographic hashes used by data-centric designs. We draw these parallels intentionally so that IBRs can support similar integrity and self-certifying properties as data-oriented designs, but at the information level. We show that it is possible to adapt the rich work on generating fingerprints for similarity identification in the context of images and videos (e.g., [9, 24, 3, 20]) for generating suitable IBRs.

To summarize, the key contributions of our paper are: (i) arguing that Web and content links must be information bound; (ii) outlining the challenges in creating and using IBRs for multimedia content; (iii) an exploration of the tradeoffs of various alternatives for generating and using IBRs, and (iv) identifying that it is possible to adapt multimedia fingerprinting algorithms to generate IBRs.

2. MOTIVATING SCENARIOS

Our goal is to enable content publishers—e.g., users writing blogs, tweets, social networking posts—to *easily* reach a wide target audience with diverse operating conditions (e.g., network bandwidths and loss rates) and diverse browsing devices, each of which may support a limited collection of file formats. We begin with two motivating scenarios and highlight the types of references needed to support them.

Multimedia in blogs and social media: There is a significant amount of user-generated content today in blogs, social networking posts, Twitter feeds, and community portals like Digg. A large fraction of this content either links to or embeds multimedia information.

When blog authors create multimedia content, they post them to a hosting site (e.g., bing.com/video) and provide a URL to the content (e.g. http://bing.com/Vid_High.flv). Unfortunately, most popular hosting sites don't support all file formats (e.g., bing.com/video is not iPhone friendly) and most offer content in a limited number of resolutions (typically 1 or 2). As shown in Figure 1, altruistic users may generate alternate versions of the content and post other resolutions (e.g. http://bing.com/Vid_Low.flv), or other formats (e.g., http://iphonefriendly.com/Vid_Low.mpg). For example, comments on sites like Digg and Slashdot often provide alternative links when the original source is inaccessible due to format issues or overload.

However, this poses a few challenges: (1) Consumers have to manually sift through comments to find the appropriate content link; (2) User comments may not be informative–e.g., a consumer cannot immediately tell if a low-res version is also iPhone friendly; and (3) Consumers have no way to check the validity of third party links and these links may lead to malicious content. Because of these issues, many users (especially those on their mobile devices) cannot access the video on their "first click" and simply move on.

Crowdsourcing for situational awareness: Community portals such as Ushahidi [2] allow users in crisis-affected places to generate image and video feeds to provide situational awareness [25, 5]. While such portals are immensely valuable, Fall et al [14] point out two key limitations: (1) users may not have uninterrupted Internet access, and (2) the network quality or bandwidth availability may be variable.

In these settings, users are often willing to contribute their own resources (e.g., laptops) to store and host content that can be viewed by others. Unfortunately, consumers of content are faced with a few challenges in making use of these resources. First, users may not trust third party sources of content. Second, users may need content in a format or fidelity different from that available at a specific peer.

In both scenarios, we want to lighten the burden of ensuring ease of access. We also want to provide mechanisms so that publishers and consumers can seamlessly integrate third-parties without worrying that their intent is being misrepresented. To this end, we need content links that can: 1) decouple content delivery from specific hosting services and support alternative transport techniques (e.g., DTN [11], P2P), (2) support the ability to reference content at different fidelities in a consistent way, and (3) provide a way to *bind* the publisher's intent to the content so that users can verify that the content has not been modified.

We envision a system (see Figure 1) in which to add a post to a blog or social networking site, a publisher creates an IBR for the content. Blog posts use URLs like IBR://12-27918-9274 instead of http://bing.com/Vid_High.flv. Publishers and hosting services can advertise IBRs that they can deliver in a presentation format.

When consumers want to access the content, they use the IBR as a handle to an *information resolution* service. This maps the IBR to suitable network locations and encodings. Thus, it can integrate third-parties; e.g., an iPhone-friendly service, and it allows users to choose suitable locations (e.g., nearby caches). User applications can then *verify* that the rendering matches the IBR given by the publisher.

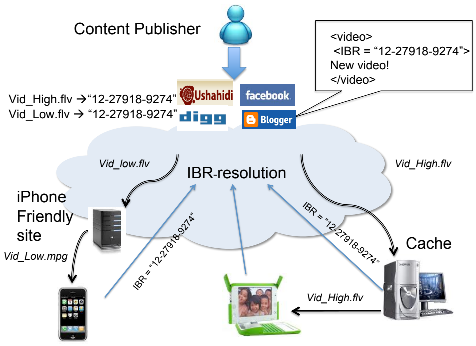


Figure 1: Example of how Information-Bound references can enable high availability, integrate third-party services, and provide dynamic adaptation

3. SOME OPTIONS

We now argue that a new approach is needed to provide IBRs. We describe a variety of existing approaches both to providing references and to hosting content, and argue how they fail to support the above scenarios. For simplicity, we focus on the first scenario.

Existing URLs + Hosting-side solutions: Several hosting services such as Hulu, NetFlix and YouTube are moving toward automatically tailoring the format and bitrate, depending on the user's device and bandwidth capabilities. YouTube already allows users to play a video in a few different resolutions and formats. Technologies such as SmoothHD are being use to enable smooth adaptation of streaming rates. New standards (e.g., HTML5) are being proposed to simplify content negotiation between hosting sites and client browsers to determine the right type of content to serve (i.e., both the encoding format and the bitrate). These approaches are insufficient on at least two fronts: First, they constrain consumers to specific formats and resolutions, i.e., those supported at the hosting service. Second, they cannot easily accommodate third parties. For instance, a hosting site may want to downgrade the resolution of a video due to bandwidth limitations, but a consumer may be able to obtain a higher quality version of the media from a different source and potentially in a different format (similar to the first example in Section 2). These approaches do not allow consumers to identify and use such alternative sources.

Data-centric and semantic-free referencing: Data-centric architectures use a cryptographic hash of the data of an object as the reference [29, 28], and a resolution system (e.g., DHTs [27]) to locate suitable sources for the object. Being *intrinsically bound* to the data object, the hash also ensures integrity: a third-party user cannot create another object that maps to the same hash value. Also, the hashes can be generated and verified in a decentralized fashion.

Because the cryptographic algorithms operate at the byte-level, references for different formats of the same information will be different. Thus, data-centric approaches allow integration of third parties in the-user generated content example, but only to a limited extent.

Requirement	URLs	Strings + Search	Data Centric	IBR
Location-independence	No	No	Yes	Yes
Encoding-invariance	Limited	Yes	No	Yes
Intrinsic binding	Limited	No	Limited	Yes
Decentralized operation	No	No	Yes	Yes

Table 1: Qualitative comparison of different proposals to support increased availability, dynamic adaptation, and the ability to integrate third-party services

The SFR system [17] has similar goals to our work. In SFR, the reference to an object is free from semantics of the domain, provider, or organization where the object is located (e.g., SFRs are of the form http://0xab329f/Vid.avi). Unfortunately, these URLs are not encoding-invariant and are, thus, unsuitable for the scenarios we target.

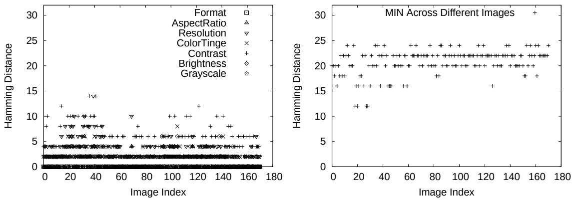
String-based references + Search: Several proposals have made a case for "intuitive" references to better capture publisher and consumer intent [6, 10, 30]. These depend on an indirection infrastructure to map these high-level tags to network locations using search engines or other resolution systems. For example, the content publisher would provide hints to express her intent (e.g., *Star Wars Trailer*) and the infrastructure (e.g., Google) would map it into a video file meeting the intent. These approaches delay the binding between the publishers' intent and the specific location or formats. They enable consumers to download content from any source and in any format. Unfortunately, using human readable names can lead to *contention* [17] and requires users to agree to a uniform naming convention and/or hierarchy. Further, publishers and consumers are now required to implicitly trust the search infrastructure to satisfy their intent.

4. INFORMATION-BOUND REFERENCES

We formally specify four requirements for IBRs:

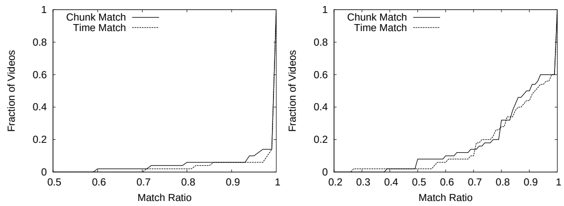
- **Location independence:** IBRs should enable anyone to serve the content pointed by the reference on behalf of the content publisher.
- **Encoding invariance:** The same content in any format, encoding, or resolution should have the same IBR.
- **Intrinsic binding:** IBRs should be *unique*; different pieces of content should have the same IBR. IBR generation should be *one-way*; it should be hard to generate fake content with a particular IBR.
- **Decentralized operation:** Generating and verifying such references should have little or no dependence on a central authority.

As Table 1 shows, prior proposals fail to meet one or more of these requirements. We also saw from these prior designs that references that are generated manually lead to contention and cannot be verified in a decentralized environment. Analogous to data-centric references that use a cryptographic hash derived from a data object, IBRs must be *algorithmically* derived from the underlying information contained in content. In the next section, we examine possible algorithms for generating IBRs for multimedia content.



(a) Across transforms (b) Across distinct images

Figure 2: Hamming distances of image IBRs



(a) Format change (b) Bitrate change

Figure 3: Distribution of match ratio (time/# chunks) w.r.t transformed versions of the video

5. ALGORITHMS FOR IBR

Multimedia fingerprinting is used in a variety of applications today such as duplicate detection (e.g., [24]), finding copyright violations (e.g., [13]), and song identification. Such fingerprinting algorithms, when designed carefully, are largely invariant across common transformations (e.g., rescaling, changing formats, changing bitrate etc.). This property allows them to be used as IBRs.

We now describe how we can extend state-of-the-art fingerprinting algorithms for generating IBRs.

Images: There are three classes of image fingerprints:

1. Algorithms that leverage spatial structure (e.g.,Gist [20]) to mimic how the human eye recognizes objects.
2. Using color distributions (e.g., [19]) that look at the distribution of R, G, and B values in an image.
3. Frequency domain analysis (e.g., [9, 3]) via Fourier or Discrete Cosine (DCT) transforms.

However, we found that the first two classes of algorithms are too coarse-grained and are subject to simple pollution attacks. For example, the spatial techniques do not distinguish grayscale and color versions; a white-black strip will have the same color distribution as an image with an equal number of black and white dots. The frequency domain techniques are more robust to such pollution attacks. As a starting point, we use the 64-bit fingerprint proposed by Coskun and Sankur [9] as our $ImgIBR$. (The high-level idea is to extract the first 64 DCT components and quantize these to 0/1 if they are higher/lower than their median.)

Figure 2 compares Hamming distance between $ImgIBR$s for transformed versions of the same image and between $ImgIBR$s for distinct images from a large image dataset [1]. The result highlights three key properties. First, the fingerprints are very similar (i.e., low Hamming distance) across transformations meaning that they provide *encoding invariance*. Second, they are dissimilar (i.e., high Hamming distance) for distinct images and, thus, provide the uniqueness part of *intrinsic binding*. Third, the Hamming distance between fingerprints are non-zero for some transformations. In other words, the IBRs generated across different transformations, while being very similar, are *not exact matches*. This becomes critical in using the IBRs in practice. We discuss the implications in the next section.

However, we do need to pick a suitable Hamming distance threshold to balance false positives (i.e., giving two different images the same IBR) vs. false negative rate (i.e., giving transformed versions of the same image different IBRs). For example, with a threshold distance of 11, we can ensure zero false positives and a 0.4% false negative rate. While this threshold is suitable for most common transformations, the robustness of such fingerprints to sophisticated *adversaries* trying to create bogus content with a specific IBR is an open question. We revisit this issue in Section 7.

Videos: A video can essentially be viewed as a sequence of images or *frames*. A simple idea is to concatenate the $ImgIBR$s for each frame and use it as the $VidIBR$. This can make the $VidIBR$s pretty large and also expensive to generate/verify.

A more efficient alternative is to choose distinct frames, logically *chunk* the video and then use the image IBRs of the boundary frames. That is, if $Video = C_1||\ldots||C_n$, we compute $VidIBR(C_i) = ImgIBR((C_i^{start}))||ImgIBR((C_i^{end}))$, where C denotes a chunk, $start$, end refer to the first and last frames in a chunk, and $||$ denotes concatenation.

Analogous to proposals for robust data-aware chunking [15, 23], we want the chunk boundaries to be derived from the underlying content, rather than imposing artificial boundaries (e.g., bytes or time durations). This will be robust to time-shifts (e.g., credits missing) and minor edits. Again, there is a rich literature on scene detection [8] we can use. These identify *keyframes* where the scene changes substantially; for example, measured in terms of signal energy or the color histogram [16].

We use a controlled dataset of 50 videos (trailers from YouTube) and apply two transforms: changing format and bitrate. For each video, we measure the *match ratio* in terms of the number of chunks and total duration of match between the source and transformed version in Figure 3. We see that format changes have minimal impact; the match rate is $\geq 95\%$ for more than 95% of the videos, both in terms of time and number of chunks. Similarly, in Figure 3(b), the match rates exceed 80% (time and # chunks) for more than 80% of the videos with the bitrate transform. (There were no false positives).

Audio: There are similar techniques for creating IBRs for audio content [21]. We omit them for brevity.

In the next two sections, we show how to use the above IBRs in practice and also highlight the main challenges involved.

6. IBR RESOLUTION

The previous section discussed algorithms for generating IBRs. To use IBRs in practice, we need an effective *resolution service* that can map the IBR given by a content publisher to specific network locations that have the desired content in a given format.

We foresee at least two key challenges in building such a resolution service: (1) embedding content-negotiation capabilities into the resolution service to allow consumers to optimally satisfy their requirements and (2) developing efficient algorithms and system designs that can naturally accommodate the fuzzy nature of IBRs. Thus, resolution must employ more complicated mechanisms than exact string matching schemes that are a common choice in other service of similar kind (e.g., DNS).

We now describe two possible resolution alternatives. The first solution depends on a logically centralized indirection service and it is explicitly designed to be backwards compatible with existing HTML/HTTP standards. Thus, it can enable publishers and consumers, such as those in the first scenario in Section 2, to immediately benefit from the power of IBRs. The latter is more suited for more challenged environments, where intermittent network access to such a service may not be available.

A URL redirection service: This provides the IBR as a link to a redirection service (see Figure 4). That is, the publisher embeds the multimedia content by providing a link to www.ibr_resolver.com/12-27918; uploads the video to a hosting service (e.g., www.bing.com/newvideo) and adds a mapping between IBR =$12 - 27918$ and the hosting site at www.ibr_resolver.com. Other users can then augment the mapping for this IBR (e.g., providing an iPhone friendly version at iphone.dailymotion.com/alt_video by adding an entry at www.ibr_resolver.com. In addition to mapping the IBR to a location, the users specify the presentation attributes such as format and bitrate to enable content consumers to select an appropriate alternative.

For the content negotiation step, we envision two alternatives. First, consumer applications can explicitly provide their requirements; that is, the browser or a browser plugin will specify the desired bitrate and supported formats. For example, a laptop's browser plugin would send its capability information along with the request to www.ibr_resolver.com. Alternatively, to support legacy clients, the service can automatically decide suitable presentation formats, e.g., by looking at user agent in the HTTP requests. For example, it can use the browser string to infer that the request is from an iPhone and thus redirect it to the iPhone-friendly site. Such URL redirection services are commonly used in blogs and Twitter feeds, albeit for URL-shortening (e.g., tinyurl.com, bit.ly). They already handle large request rates ($\approx$ 2-3 billion queries per month) and so we believe scalability will not be an issue. Optionally, the redirection service can return a set of potential matches so that users can choose what they perceive to be the best fit; e.g., they can choose preferred hosting providers.

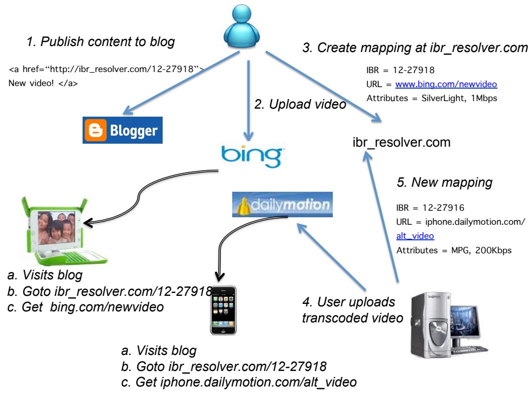


Figure 4: Example showing how a URL redirection service can resolve IBR to actual network locations

As we saw in the previous section, the IBRs generated across different encodings might differ slightly. For example, in Figure 4, the IBR for the iPhone-friendly version is $12 - 27916$, whereas the original IBR in the blog entry was $12 - 27918$. A simple exact-match interface would not suffice for this resolution service. However, we believe that techniques such as locality sensitive hashing [12] can be used to build the mapping system at the resolution service. In addition, we can leverage traditional database capabilities to provide the range-query like functionality needed to support content negotiation and specify constraints on the presentation attributes (e.g., to specify encoding rate < 400 Kbps).

A decentralized lookup service: In scenarios such as the situational awareness application from Section 2, intermittent connectivity may make such a URL redirection service inaccessible. In this case, we can extend techniques for DHT-based publish-subscribe systems [22, 28] to find nearby or suitable peers that have content with a specific IBR. Because IBRs may differ slightly across encodings, we need to extend such DHT-architectures to exploit locality in the IBR space [4, 26, 7].

7. SUMMARY AND DISCUSSION

In this paper, we made a case for Information-Bound References (IBRs) to support easy multimedia dissemination and access. We presented compelling examples of scenarios where such references make multimedia access remarkably more flexible than the existing URL-based approaches. We also argued why other strawman schemes, such as data centric references, fail to benefit these scenarios. We argued that it may be possible to leverage algorithms from multimedia fingerprinting as the basis for IBRs.

We conclude the paper by describing additional interesting advantages that using IBRs offers. We also highlight key challenges pertaining to IBRs that need to be addressed in the near future.

7.1 Opportunities

Supporting evolution in media formats: Media formats continue to evolve over time, both as a result of the development of more efficient encoding/compression algorithms and as the need for higher fidelity formats arise with better rendering technology. With today's URLs, a publisher wishing to upgrade old multimedia content to new formats would be forced to sift through all previously published web pages that contain references to a given multimedia content, and either modify the reference therein, or add new references. Using IBR, publishers have a much simpler task–they add a new mapping at the resolution service.

Video caching: IBRs can be employed in designing interesting new multimedia caches. In traditional Web or P2P caches, the requested URL or content hash is used as a handle to identify if an item is available in cache. Similarly, the IBR requested by a consumer can be employed to identify if alternate versions of a media file are available locally. Based on knowledge of the encoding and resolution that the consumer is willing to accept (this can be derived from browser strings or the consumer can explicitly provide these parameters), the cache can serve the most optimally-suited version. Since these caches exploit similarities across formats, they can be more effective at reducing network load from repetitive transmissions than traditional URL- or data-centric caches.

7.2 Challenges

Content pollution: Designing IBR generation algorithms that are one-way and, thus, robust to adversarial attacks (e.g., trying to create inappropriate content that will have a specific IBR that matches some popular content) is an open question. Until such algorithms are developed, we can rely on out-of-band mechanisms to offer enhanced integrity guarantees. For instance, a consumer can simply trust only IBRs from well-known providers, peers or hosting sites. More generally, we could also leverage a trusted third party to attest to the binding between content in a given format and its IBR; subsequently we only need a cryptographic hash to ensure information integrity. Finally, even if we do not trust the sources, we can use reputation systems like Credence [31] to enhance the integrity guarantees obtained from using IBRs.

Handling evolution in IBR algorithms: Multimedia fingerprinting is an active area of research today. It is likely that the algorithms used to generate IBRs will change over time. To support this, one option is to add versioning and algorithm information to the metadata along with the IBR stored at the resolution service. This would make it possible to support multiple fingerprinting techniques. However, there are some open questions to consider, such as: How and when are new IBRs computed? To what degree should the system support translation between names? What if videos needs to be re-chunked for the new algorithms? We leave an exploration of this important issue for future work.

Acknowledgments

This work was support in part by NSF grants CNS-1040757, CNS-1032226, CNS-0905134 and CNS-0746531, by CyLab at CMU under Federal Grant W911NF0910273 and by a gift from Cisco. Ashok Anand's research is supported by a Google PhD Fellowship.

8. REFERENCES

[1] http://www.cs.washington.edu/research/imagedatabase/.

[2] Crowdsourcing crisis information. `http://www.ushahidi.com`.

[3] The open source perceptual hash library. `http://phash.org`.

[4] N. J. A. Harvey, M. B. Jones, S. Saroiu, M. Theimer, and A. Wolman. SkipNet: a scalable overlay network with practical locality properties. In *USITS*, 2003.

[5] A. Hughes, and L. Palen. Twitter Adoption and Use in Mass Convergence and Emergency Events. In *Proceedings of ISCRAM*, 2009.

[6] W. Adjie-Winoto, E. Schwartz, H. Balakrishnan, and J. Lilley. The design and implementation of an intentional naming system. In *SOSP*, 1999.

[7] Q. A. Bharambe and A. R. Bharambe. Mercury: Supporting scalable multi-attribute range. In *In SIGCOMM*, pages 353–366, 2004.

[8] J. Boreczky, , J. S. Boreczky, and L. A. Rowe. Comparison of video shot boundary detection techniques. In *In Storage and Retrieval for Still Image and Video Databases IV, No. SPIE 2664*, pages 170–179, 1996.

[9] B. Coskun and B. Sankur. Robust video hash extraction. In *Proc. EUSIPCO 04*, 2004.

[10] D. Trossen, M. Sarela, K. Sollins. Arguments for an Information-Centric Internetworking Architecture. *ACM SIGCOMM CCR*, Apr. 2010.

[11] K. Fall. A delay-tolerant network architecture for challenged internets. In *Proc SIGCOMM*, 2003.

[12] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *Proc. VLDB*, 1999.

[13] P. Indyk and N. Shivakumar. Finding pirated video sequences on the internet. Stanford Infolab Technical Report, Feb. 1999.

[14] K. Fall, G. Iannaccone, J. Kannan, F. Silveira, and N. Taft. A Disruption-Tolerant Architecture for Secure and Efficient Disaster Response Communications. In *Proceedings of ISCRAM*, 2010.

[15] R. M. Karp and M. O. Rabin. Efficient randomized pattern-matching algorithms. *IBM J. Res. Dev.*, 31(2):249–260, 1987.

[16] R. Kasturi, S. H. Strayer, U. Gargi, and S. Antani. An evaluation of color histogram based methods in video indexing. Technical report, Penn State University, 1996.

[17] M. Walfish, H. Balakrishnan, and S. Shenker. Untangling the web from DNS. In *Proceedings of NSDI*, 2004.

[18] Michael Freedman. Experiences with CoralCDN: A Five-Year Operational View. In *Proc NSDI*, 2010.

[19] C. L. Novak and S. A. Shafer. Anatomy of a color histogram. In *CVPR*, pages 599–605, 1992.

[20] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. *Progress in brain research*, 2006.

[21] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A Review of Audio Fingerprinting. *The Journal of VLSI Signal Processing*, 41(3), 2005.

[22] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec. The many faces of publish/subscribe. *ACM Computing Surveys*, 35(2), June 2003.

[23] H. Pucha, D. G. Andersen, and M. Kaminsky. Exploiting similarity for multi-source downloads using file handprints. In *Proc. of NSDI*, 2007.

[24] S.-C. Cheung and A. Zakhor. Estimation of web video multiplicity. In *Proc. SPIE Internet Imaging*, 2000.

[25] S. Liu, L. Palen, J. Sutton, A. Hughes and S. Vieweg. In Search of the Bigger Picture: The Emergent Role of On-Line Photo-Sharing in Times of Disaster. In *Proceedings of ISCRAM*, 2008.

[26] I. Stoica, D. Adkins, S. Zhuang, S. Shenker, and S. Surana. Internet Indirection Infrastructure. In *Proceedings of ACM SIGCOMM*, 2002.

[27] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for Internet applications. 2001.

[28] T. Koponen et al. A Data-Oriented (and Beyond) Network Architecture. In *Proc. SIGCOMM*, 2007.

[29] N. Tolia, M. Kaminsky, D. G. Andersen, and S. Patil. An architecture for Internet data transfer. In *Proc. NSDI*, 2006.

[30] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs, and R. Braynard. Networking Named Content. In *Proc. CoNEXT*, 2009.

[31] K. Walsh and E. G. Sirer. Experience With A Distributed Object Reputation System for Peer-to-Peer Filesharing. In *Proc. NSDI*, 2006.