Multimed Tools Appl (2014) 70:1069–1098
DOI 10.1007/s11042-012-1103-3

Community and trust-aware fake media detection

**Khaled Ahmed Nagi Rashed · Dominik Renzel ·
Ralf Klamma · Matthias Jarke**

Published online: 3 May 2012

Abstract Nowadays, it becomes increasingly difficult to find reliable multimedia content in the Web 2.0. Open decentralized networks (on the Web) are populated with lots of unauthenticated agents providing fake multimedia. Conventional automatic detection and authentication approaches lack scalability and the ability to capture media semantics by means of forgery. Using them in online scenarios is computationally expensive. Thus, our aim was to develop a trust-aware community approach to facilitate fake media detection. In this paper, we present our approach and highlight four important outcomes. First, a Media Quality Profile (MQP) is proposed for multimedia evaluation and semantic classification with one substantial part on estimating media authenticity based on trust-aware community ratings. Second, we employ the concept of serious gaming in our collaborative fake media detection approach overcoming the cold-start problem and providing sufficient data powering our Media Quality Profile. Third, we identify the notion of confidence, trust, distrust and their dynamics as necessary refinements of existing trust models. Finally, we improve the precision of trust-aware aggregated media authenticity ratings by introducing a trust inference algorithm for yet unknown sources uploading and rating media.

K. A. N. Rashed (✉) · D. Renzel · R. Klamma · M. Jarke
Computer Science 5—Information Systems & Databases,
RWTH Aachen University,
Ahornstr. 55, 52056 Aachen, Germany
e-mail: rashed@dbis.rwth-aachen.de

D. Renzel
e-mail: renzel@dbis.rwth-aachen.de

R. Klamma
e-mail: klamma@dbis.rwth-aachen.de

M. Jarke
e-mail: jarke@dbis.rwth-aachen.de

Keywords Community-based approach · Trust-awareness · Multimedia Quality Profile · Collaborative fake media detection · Multimedia reputation · Semantics for faked multimedia

1 Introduction

Today's Web 2.0 platforms experience an explosive growth of multimedia content. Such platforms provide services for sharing and distributing user generated multimedia content. Most of the media produced and shared are only valuable for a limited circle of users and do not create any awareness beyond friends, family and fools. However, typical misbehavior on the Internet such as cyber-bullying [30, 35] is transferred now to multimedia materials, supported by new computational tools such as face recognition as implemented in Facebook. Due to the nature of such platforms, an agent can often publish media without going through any authenticity checking mechanism. These systems are often characterized by anonymity and dynamics which make them vulnerable to misbehaving or even criminal users. These users not only spread fake multimedia, but also facilitate massive compromised evaluation of fake media realized with fake accounts and bot farms acting as automated agents.

Misbehaviour can be emergent in many ways by using these open environments. In some cases, misbehaving users manipulate images and distribute them as authentic or publish real ones with false metadata. Some users may act with malicious intent to throttle the expectations of users to make them lose confidence in the system. Other users may just want to maximize their gains (reputation, expertise, etc.) and/or maximize/minimize the reputation of particular media by increasing/decreasing their ratings for that media. Users can also act maliciously by negatively influencing other participating users' rankings and trustworthiness.

Web 2.0 platforms are now used in professional settings like media press agencies dealing with multimedia materials. An increasing number of media distributors relies on contributions from amateur reporters producing authentic materials on the spot, e.g in cases of natural disasters or political disturbances. However, even with mobile devices it is easy to forge media on the spot of capturing and publishing them. Thus, it is increasingly harder to determine the originality and quality of delivered media, especially under the constant pressure to be first on the news market. Especially in highly sensitive areas such as politics, security and business the distribution of fake multimedia can destroy the reputation of whole organizations, institutions, even nations. Well known cases are faked media in war propaganda like in the Iraq war and faked historic pictures neglecting the existence or presence of communist party members in Russia or China.

By *fake* we refer to intentional media manipulation by changing its content and/or context to bring changes that influence opinions, facts and representations of real world events. Fake can be achieved with a multitude of different manipulation techniques [8]:

1. Adding details into media by inserting regions or object from the same image or from another images and adapting them to fit into the entire media environment (cf. Fig. 1a).
2. Deletion of media details, by removing scene elements (regions) and replacing them by others (cf. Fig. 1b).

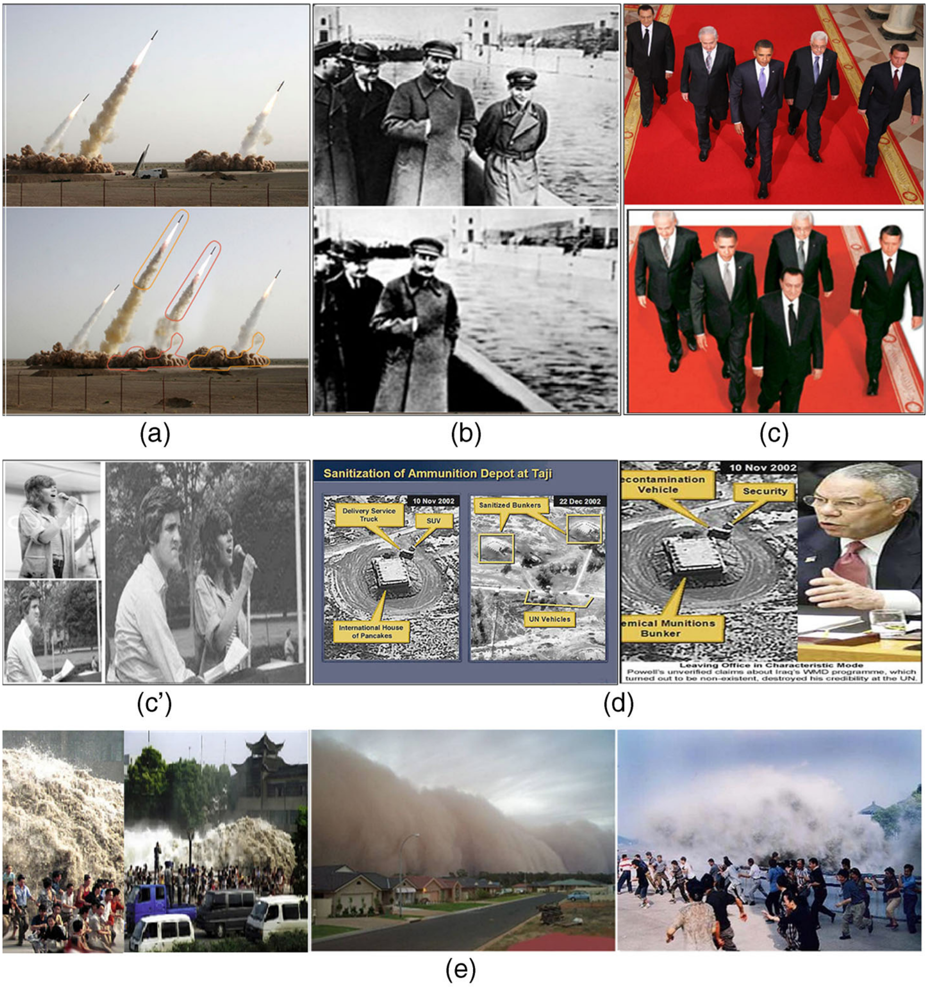

(a)
(b)
(c)
(c')
Sanitization of Ammunition Depot at Taji
10 Nov 2002
Delivery Service Truck
SUV
International House of Pancakes
22 Dec 2002
Sanitized Bunkers
UN Vehicles
10 Nov 2002
econtamination Vehicle
Security
emical Munitions Bunker
Leaving Office in Characteristic Mode
Powell's unverified claims about Iraq's WMD programme, which turned out to be non-existent, destroyed his credibility at the UN.
(d)
(e)

Fig. 1 Fake media examples: **a** addition regions, **b** deletion regions, **c** montage, and **d**, **e** metadata forgery or false captioning

3. Generating montages by combining separate images, also called composition and splicing in literature (cf. Fig. 1b, c).
4. False captioning as mislabeling or inaccurately dating media without necessarily modifying their content. False captioning can be defined as the falsely captioned media differs from other groups of fake photos in that, although the media has not been altered, the context of what the media purportedly conveys is simply falsified (cf. Fig. 1d, e). False captioning includes metadata counterfeiting such as wrong names, events and places or other metadata forgeries, which distort EXIF technical information such as GPS data. The pictures in Fig. 1e are real pictures published to demonstrate other events claiming to be what they are not. The pictures in Fig. 1e (left and right) were posted on the Web claiming to show the impact of the 2004 Boxing Day Tsunami. However, they actually show tourists

gathered to watch the results of a tidal bore on the Qian Tang Jian River in China 2002. Metadata of the pictures in the middle claim to show a street in Sumatra just before it was inundated by a gigantic wave from the tsunami that hit Indonesia in 2004. In fact, the pictures show a massive dust storm that hit the town of Griffith, NSW Australia on November 13, 2002.

5. Incorrect classification, ordering and placement in presentations of media collections.
6. A combination of two or more operations above-mentioned.

Media alteration by means of changing media semantics with the aim to deceive the public is center of our concerns in the context of this work. We consider above listed operations, while another types of media transformation where there is no local manipulation or pixel combination, i.e. rotation, smoothing, luminance change, cropping, compression and additive noise are accounted as versions of the original ones.

We argue that fake multimedia understanding is community dependent. Multimedia can be altered for many reasons: fraud, greed, malice, humor, profit, deception, education and to sway public opinion, to rewrite history and to show discontent [8]. Digital photographs for instance in the news industry are often adjusted for reasons of aesthetics by changing the contrast here, a color-alteration there. But they can also be altered with the aim to deceive public.

In the scope of this work, we are not interested in the analog multimedia representation. Instead we deal with digitized multimedia representation. Digitization of media opens the door to their easy manipulation. According to [49], the principles of media digitalization have made the process of media manipulation, organization, sharing, transferring, and remixing easier and increasingly automated. Furthermore, the recent advances in digital multimedia technologies such as acquisition, processing, ease of hosting, transmitting and migrating digital media over the Web have complicated the important task of determining media authenticity.

Automatic and cryptographic techniques [17, 31, 40, 65] are used to ensure or protect media integrity, but exhibit problems with context and semantics. Automatic approaches are not able to detect the iconographic meaning of a picture determining the importance for the spectator. In our previous contribution we presented a community-based approach for avoiding fake media distribution and untrusted media evaluation [63]. We faced the problem of trustworthiness of agents involved in this approach. Furthermore, the cold start problem posed the challenge of no media and no users being available for evaluation. The elicitation of contributions from users is problematic [64] because of a lack of incentives for giving feedback or actively contributing content.

In this article, we describe a new approach for facilitating faked media management and detection in community-based social networks. Our contributions are as follows.

- We developed a new Media Quality Profile for semantic multimedia classification. To the best of our knowledge, our MQP is the first trust-aware Media Quality Profile that includes multimedia reputation metadata powered by trust and distrust among people and confidence in the system in its computation.

- We developed a serious gaming platform to address the fake media detection community cold-start problem. We identified the cold-start problem in our community-based fake media detection system, and thus designed a game platform to engage participants to provide media authenticity ratings, while they enjoy the game and receive incentives to keep playing.
- We identify the notion of confidence, trust, distrust and their dynamics as necessary refinements of existing models.
- We improve the precision in computing trust-aware aggregated media authenticity ratings by introducing a trust inference algorithm for yet unknown sources.

The rest of this article is structured as follows. In Section 2 we first describe our Multimedia Quality Profile with a focus on the inclusion of trust-aware media reputation information. In the following sections, we discuss the components necessary to power our MQP with sufficient data and computing techniques. In Section 3 we describe how to initially solve the cold-start problem by introducing a community platform for playing simple media authenticity rating games and thus to guarantee sufficient multimedia content accompanied by rating data. In Section 4 we discuss the characteristics and challenges of trust management among participants in such systems and present a model, an implementation, and an evaluation of our approach to compute trust-aware aggregate media authenticity ratings as proxy measures to media reputation in our MQP. Finally, we conclude and provide an outlook to future work in Section 5.

2 The Multimedia Quality Profile MQP

Multimedia are now used in wide areas and distributed in different ways. Multimedia authenticity is becoming increasingly important. While the vast majority of trivial multimedia alterations have trivial consequences and are thus acceptable, some sorts of multimedia forgery can cause serious problems. Automatic detetion techniques have their limitations. Therefore, we can only rely on the Web 2.0 and community methods to judging media authenticity. Media quality in terms of authenticity information must be captured and presented to community members in a meaningful way to become useful. To help managing fake media detection, we need measures and mechanisms to support judging media authenticity. Therefore, we propose our MQP to help community members in their decisions about multimedia quality. However, designing a Multimedia Quality Profile requites a comprehensive understanding of multimedia distribution, semantics, evaluation problems and dynamics within community systems.

This section presents a detailed description of our proposed Multimedia Quality Profile (MQP). The aim is to create a data model including both automatically computed and community-generated information. The media can be described in many different ways regarding for example their content from either a technical or a semantic point of view, its technical attributes, and its quality. Our MQP consists of combinations of metadata drawn from various parts of our fake multimedia management and detection system [61], namely media annotation, visual feature extraction, and community and trust-aware media authenticity rating management. MQP is based on a few main components, which are acquired either automatically by a device or manually by communities of people, and which act as combinable building blocks to support different kinds of use cases (cf. Fig. 2):

1. *Exchangeable Image File Format (EXIF) descriptions:*
   EXIF [77] is a standard that specifies the formats for images, sound, and ancillary tags used by digital cameras, scanners and other systems handling image and sound files recorded with digital cameras.
2. *MPEG-7-based descriptions:*
   MPEG-7-based descriptions [47] in our MQP include multimedia content and semantic descriptors. MPEG-7 multimedia content descriptions may include information describing the creation and production processes of the content, information pertaining to the usage, storage features of the content and information about the interaction of the user with the media content. It can be on a global scope (i.e. describing only metadata related to a complete media item, such as title and production information) or related to spatial, temporal and spatiotemporal segments of the content. MPEG-7 semantic base types are used for the description of semantic entities present in the multimedia content, i.e. a set of semantic concepts *agents, objects, time points, places, events, concepts, and states* to cover the semantic space queries. Events can be perceived as occasions upon which something happens. Agents, time, objects and places can further describe such occasions. These entities can have properties and states and are interrelated among themselves.
   Furthermore, MPEG-7 allows for the storage of low-level features. We should note that these features do not provide comprehensive information about a medium being authentic, but they support to the process of community decision in different ways: *Similarity Search*. The comparison of low-level features of different media enables similarity search and content-based media retrieval. Given a yet unknown picture a mediator can first apply content-based image retrieval on available multimedia repositories to find out whether similar images

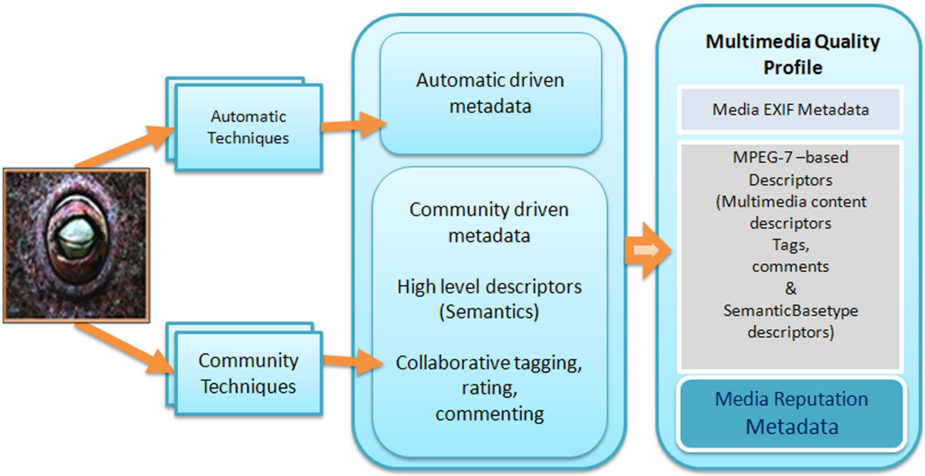

Automatic Techniques
Community Techniques
Automatic driven metadata
Community driven metadata
High level descriptors (Semantics)
Collaborative tagging, rating, commenting
Multimedia Quality Profile
Media EXIF Metadata
MPEG-7 –based Descriptors (Multimedia content descriptors Tags, comments & SemanticBasetype descriptors)
Media Reputation Metadata

Fig. 2 Media metadata

are present. In case of finding similar pictures the trust towards both given and reference image source is lowered and distrust raised. In our work we make use of MPEG-7 color and texture feature descriptors [48], both of which are based on the HSV color space and denoted by histograms. Among many useful MPEG-7 visual descriptors, we consider color layout, scalable color, dominant color, edge histogram and color structure to create an image signature.
Identification of media Alterations. In the case of existing similar media visual signature tools [7, 32] can be applied to affirm the presence or absence of media alterations. Image signatures are extremely robust tools for identifying images with alternations and editing effects.
In our system we implement several Web services enabling the community-aware creation, management and usage of MPEG-7 data. In particular we employ services for multimedia content and semantic base type and multimedia content descriptions including functionality such as low-level feature extraction, similarity matching, and content based image retrieval [37, 74]. In particular, we use Caliph & Emir [45, 46] for the extraction of MPEG-7 visual descriptors and Exif data.

3. *Media reputation information:*
One of the most important metrics we promote in our fake media detection approach are the trustworthiness of media sources and evaluators, and in consequence media reputation. Due to the risk of unfair media evaluation, using the mere average over all ratings as a quality measure may trend towards superficial evaluation or even worse to attackability. Therefore, the trustworthiness of the community agents providing media authenticity ratings is used as a weighting factor. We assume that if a particular community agent is trustworthy, it is likely that he will provide fair ratings. We thus propose to have some criteria or quality dimensions by asking the following fundamental questions:

   - Is the media source trustable?
   - Is there any similar version of the medium?
   - Are media authenticity ratings trustable?
   - Is a critical mass of trustable authenticity ratings reached?
   - Is the number of trusted raters acceptable?
   - Which experts have been involved in the media evaluation?
   - Is the resulting media reputation credible?

   A majority of positive answers to the above questions can be an good indicator for right decisions about media authenticity. As decision some rules can be used such as *if a medium has a reputation value smaller than a specific threshold, then it is likely to be fake; otherwise it is likely to be real*. Having a Media Quality Profile with such metadata thus provides substantial decision support for media authenticity. Moreover, it will allow us to identify and explore key risks of quality of a medium to decide its publication and distribution. To the best of our knowledge, no standardized multimedia quality profile including media reputation information exists. Thus, we propose to develop an own schema, based on the above discussions. Using an XML format for storing such a quality profile benefits from the extensibility property, since new descriptors and metrics can be added easily. RDF can also be used, which could facilitate web searches and semantic queries.

For media evaluation, the expertises and reputations of selected media evaluators or groups of experts serve as a guarantee of the media quality. Thus, collective aggregated media authenticity ratings of community members must be augmented with information about numbers, reputation and expertise of raters to come to a credible media quality indicator. In our system, we identify the following roles involved in establishing information about media authenticity:

- *Mediator:* agents that control the media quality and decide publishing or rejecting media, establishing thresholds, monitor changes
- *Normal user:* agents that contribute/add/rate media
- *Expert:* a highly trusted user and/or domain expert
- *Malicious user:* an agent that intentionally degrade or upgrade the media quality evaluation to misinterpretation it, they can exist as individual and collusive malicious evaluators.

Consequently, the media reputation included in our MQP is in general constructed through dynamic interaction among the members of the above mentioned groups contributing in different activities and playing different roles. An expert is expected to provide more reliable media authenticity ratings than someone who is new to this topic or domain. Thus, media evaluation done by experts and highly trusted users will be more valuable.

An XML schema for media reputation information We follow an attribute-based data model used extensively in data quality literature (e.g. [79]). By attributes we refer to the characteristics that affect the overall media evaluation and by metrics refer to measures for the given attributes. As it can be seen in Listing 1, an XML schema has been created that prescribes the shape of the resulting XML representation of the trust-aware media reputation part of the MQP. The main element is *Reputation*, the root which groups the data values: *Rates* and the number of trusted users (*NrOfTrustedUsers*) which will be considered when the aggregated rate will be computed. The *Rates* element stores the sequence of ratings that were registered for the media until the moment represented by the *TimeStamp* attribute. Each *Rate* element stores a rating assigned by a given user together at a given timepoint represented by the attribute *dateTime*. The aggregated rating will be computed separately for each user using the user's local trust values. In Section 4 we present how these values are computed. Those ratings considered for the aggregated media authenticity rating will be the first *NrOfTrustedUsers* ratings ordered in a descending way by the trust values the user assigned to the raters. To keep track of the way the reputation is changing in time, the previous data contained will be stored in a materialized view whenever a new MQP is created.

Monitoring media status We use a widget interface continuously displaying media reputation on a real-time basis (cf. Fig. 3).

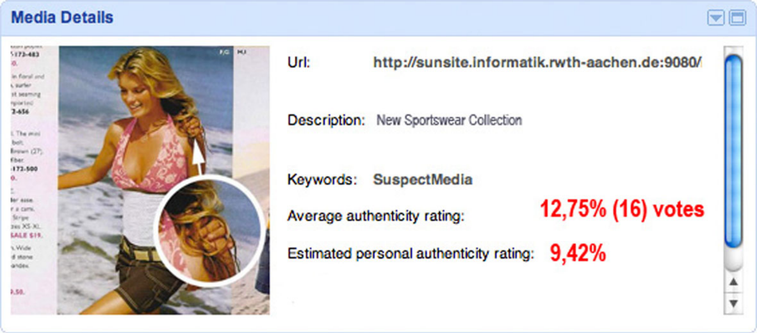
Results are currently displayed in a text format and can be augmented with the corresponding flags coding the current status of the particular medium. We should note that part of the displayed information is viewer dependent, since it depends on the trust relation the user has regarding the underlying media at a certain point in time. Once a medium has been uploaded and the evaluation has been started, a default MQP instance is generated. Then, its values are triggered after adding new

```
<?xml version="1.0" encoding="UTF-8"?>
<schema xmlns="http://www.w3.org/2001/XMLSchema"
 targetNamespace="http://www.example.org/Reputation"
 elementFormDefault="qualified">
  <element name="Reputation">
  <complexType>
   <all>
    <element name="Rates">
     <complexType mixed="True">
      <sequence>
       <element name="Rate" minOccurs="0" maxOccurs="unbounded">
        <complexType>
         <simpleContent>
          <extension base="string">
           <attribute name="authorId" type="IDREF" use="required"/>
           <attribute name="dateTime" type="dateTime" use="required"/>
           <attribute name="expertiseValue" type="double" use="required"/>
          </extension>
         </simpleContent>
        </complexType>
       </element>
      </sequence>
      <attribute name="TimeStamp" type="dateTime" use="required"/>
     </complexType>
     </element>
    <element name="NrOfTrustedUsers"></element>
   </all>
  </complexType>
 </element>
</schema>
```

Listing 1 XSD schema for the MQP media reputation part

media authenticity ratings. We describe dynamicity in terms of trust progression in Section 4.2). At a certain point in time the media quality profile can be automatically generated by profiling the evaluation history up to that point in time. We accept only one valid rating during each time interval to avoid repeated malicious authenticity ratings. If there does not exist a sufficient amount of ratings given by a sufficiently large group of trustworthy users, then no aggregated rating can be computed.

Since media reputation in terms of fakes is considered a proxy for media authenticity, we established a limit in the number of trusted users as a threshold. The confidence in the resulted media reputation value (aggregated authenticity ratings)



Url: http://sunsite.informatik.rwth-aachen.de:9080/

Description: New Sportswear Collection

Keywords: SuspectMedia

Average authenticity rating: 12,75% (16) votes

Estimated personal authenticity rating: 9,42%

Fig. 3 MQP data including trust-aware aggregated media authenticity ratings

can be identified as a scaling factor according to the number of trusted users or experts to take a decision. The final decision can take the mean of media reputation and the confidence.

We are implementing a set of Web services to generate media quality profile instances which combining automatically generated and community-driven metadata. For security considerations regarding our proposed quality profile, digital signatures can be embedded to verify its integrity and to prevent tampering.

3 Cold-start problem and media authenticity rating games

One of the main challenges in Web 2.0 communities is finding a solution to Nielsen's participation inequality (*90-9-1 rule*) [57] and the user and item *cold-start problem* [69] describing the lack of users and media in the system and/or the lack of detailed user and media profiles. Our community-based fake media detection system can only be effective if a critical mass of trustable media authenticity ratings is available. The challenges here are how to ensure the usability of the system to sustainably attract a critical mass of users especially right after the launch of the system where an insufficient amount of media and authenticity ratings has been collected.

Different approaches have been proposed to address this challenge. Kim [36] argues that the key is in the design of an online community. For starting and maintaining a successful online community he introduced nine strategies. Turner [78] proposes to import existing contents and user profiles into the system. Others provide incentives to encourage participation [10, 26] or attempt to hire users to seed their sites (cf. Amazon's Mechanical Turk). In an educational context, learners required to submit their homework via community systems are provided with participation bonus for using the system [81]. To not only regulate the quantity, but also the quality of agents' media authenticity ratings and to ensure a sustainable level of participation in a community-based fake media detection system, it is important to create participation incentives. As a motivation strategy [10], we have designed a portal for sharing media and playing a suite of media authenticity rating mini-games following the principles of [3, 36]. Our goal is to create low entry barriers and strong binding of user interest to sustainably engage as many users as possible in sharing and evaluating the authenticity of media. In the next section, we discuss our gaming approach and its use in our community-based fake media detection system to deal with the cold-start problem.

3.1 Serious games in community-based fake media detection system

Benefiting from a growth rate of social networking platforms and increasing numbers of interfaces offered by them to leverage and extend their built-in core functionality. S erious games are a proven way of leveraging the wisdom of the crowds to tackle tasks that are easy to solve by humans but cannot yet be solved by computer algorithms in an appropriate way [60]. Serious games use entertainment technology, principles and creativity to build games that carry out real world objectives. Therefore, serious games are in our context just a result of applying games and simulation technology to non-entertaining objectives [89]. Thus, they are called *Games with a Purpose (GWAP)*.

Although GWAP can be entertaining, they have been used for different purposes including educational training, health care,[1] political games,[2] military training, strategic communication such as America's Army,[3] image labeling such as the ESP Game.[4] In the Goggle Image Labeler,[5] players provide meaningful labels which in consequence improve the Web-based search as a side effect of playing the game and many other sectors of society (See an overview of serious games [76]). Following a crowdsourcing approach in the form of a game with a purpose, our games serve the main goal of motivating users to generate a high number of media authenticity ratings, yielding a large set of manually rated media to be evaluated for fakeness. This information is then used for searching and ranking experts and highly trusted users and powering the media reputation part of the MQP.

Motivating users The main challenges when creating GWAPs is motivating users to play the game while generating useful data. There are two factors which any Web community has to take into account. First, how do users interact with each other? Second, what do they expect to get in return for their involvement and contributions? Games have to include incentive schemes that engage users while delivering considerable amounts of artifacts in terms of media authenticity ratings to overcome the cold start problem. In collaborative and social networking platforms, the main motivation to play the game is the level of entertainment they provide, as well as additional incentives such as experience points, levels or achievements that may be built into the application. Asking ourselves, what could motivate people to dedicate their valuable time to rating media authenticity, we see the following options:

1. We assume people will contribute because they have intrinsic interest in multimedia and media evaluations and simply contribute because they believe that they do great work for mankind. In this regard, the social facilitation effect is considered a widely harnessed non-monetary incentive mechanism to promote increased contributions to online systems [86].
2. Providing an awarding mechanism and non-monetary incentives as source of motivation for user contribution in an authenticity rating process.
3. Providing economic incentives, by means of paying money to users to encourage them to participate and contribute in multimedia authenticity rating similar to some social network platforms that attempt to hire users their sites using Amazon's Mechanical Turk[6] or CrowdFlower.[7]
4. Setting up an authoring and rating mechanism enforcing data acquisition.

[1] www.legacyinteractive.com

[2] www.takebackillinoisgame.com

[3] www.americasarmy.com

[4] www.espgame.org

[5] images.google.com/imagelabeler/

[6] www.mturk.com/mturk/

[7] http://crowdflower.com

Furthermore, we support the effect of other proven sources of motivation that have been studied in online collaborative sites such as Wikipedia (cf. [11, 39]).

Media authenticity rating games design Our work adopts Luis von Ahn's paradigm and principles [2] for designing our mini-games.

1. Simplicity. Keep it simple.
2. Fun is a predominant user experience. Research suggests that incorporating game elements into user interfaces increases user motivation and the playfulness of work activities [3, 71]. According to [71], besides providing the right functionality and usability features, engaging users with fun features is an important design goal for game interfaces. These include attractive graphics, appealing sound, animations, etc. (cf. the *Eight Golden Rules* for designing user interfaces [72]).
3. Collective intelligence. The wisdom of crowds mechanism is adopted in our games. If many (trusted) users say that a medium is fake, then it is likely to be fake. It has been reported that groups of users perform well only under certain conditions to fulfill the requirements to tap the wisdom of crowds.
4. Massive user participation. Our games aim at massive user participation. The assumption about the effectiveness of collective intelligence by means of wisdom of the crowds is true only by reaching a critical mass of user participation and massive generated data in terms of media authenticity ratings.

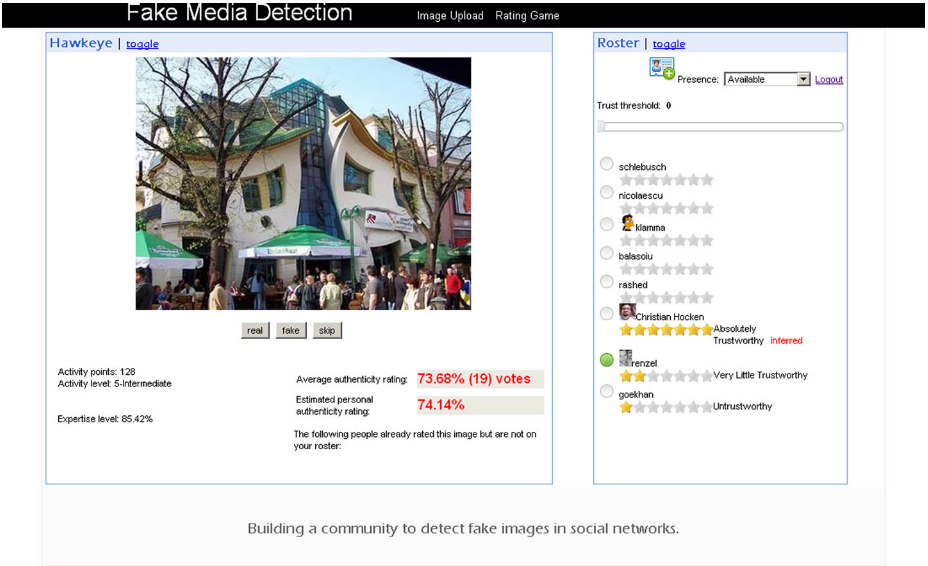
Requirements & challenges The hidden goal behind our serious gaming approach as mentioned above is to collect and acquire multimedia metadata (ratings, tags, media content). From the data produced in the games, we derive the media and user reputations later on used for generating an MQP instance. When hiding the fakemedia detection community cold-start problem and media reputation generation behind mini-games the following challenges and requirements are taken into consideration:

- Designing the games scenarios, be sure that the games conceptual design is made such that they are fun to play and achieve their main goal.
- Creating the user interface usable, attractive and aesthetically appealing.
- Aquire a large media dataset for more precise information.
- Ensure fairness of generated data. However, expecting users to provide honest and fair ratings seems to be a challenge. It is required to develop a mechanism that renders malicious behavior difficult or at least minimizes its occurrence and impact.
- Ensure scalability by balancing user distribution over time.
- Ensure sustainable contribution.
- Introduce user roles such as rater, tagger and/or media uploader.
- Provide additional incentives to sustainably play the games.

Game play & incentive scheme In this section we present a detailed description of our rating games. To regulate quantity and quality of contributed media authenticity ratings and to ensure a sustainable level of participation in our community-based fake media detection system, it is important to create participation incentives. As a motivation strategy [10], we have designed a portal for sharing media and playing a suite of media authenticity rating mini-games following the principles of [3, 36]. Our goal is to create low entry barriers and strong binding of user interest to sustainably engage as many users as possible in sharing and evaluating the authenticity of media content. We support some main ideas of gamification in terms of earning points and achievements and reaching levels. Figure 4 shows the "HawkEye" game as an example, where the player is presented a medium with yet unknown authenticity status and should simply rate it as fake or real. Of all uploaded suspect images, a random medium not yet rated is displayed to the player. The player can rate the image as fake or real. Additionally, he can request a clue that indicates what ratings have already been given by other players for that particular image. If unsure, he can always skip to the next image.

Since the purpose regarding the games is to collect as many ratings as possible, players can and should play as much and as long as they want. Other games involve the player in finding locations of forgery. Multiplayer variants are also imagineable, e.g. who rates the most pictures in a given timeframe. In the course of playing these games and providing ratings repeatedly, the player collects experience points, eventually resulting in higher levels . Additionally, players can earn and collect achievements as further symbolic incentives to continue playing. All of these common concepts present in current online games aim to bind players over a long time.

It should be noted that a scoring system for right or wrong ratings is problematic due to the possibly long delays between the player's rating and the final decision of a mediator based on the aggregated result of the collaborative rating process (cf. [63])—only if the medium is decided to be real or fake, an achievement can be added to the player's profile. Thus, we make a difference between normal *community activity points and achievements* granted for any ratings, regardless of their correctness, and special *community expert points and achievements* granted for trustable and correct ratings according to our trust update algorithm described in more detail in Section 4.



Activity points: 128
Activity level: 5-Intermediate

Expertise level: 85.42%

Average authenticity rating: 73.68% (19) votes

Estimated personal authenticity rating: 74.14%

The following people already rated this image but are not on your roster:

Trust threshold: 0

○ schlebusch
○ nicolaescu
○ klamma
○ balasoiu
○ rashed
○ Christian Hocken ★★★★★★★ Absolutely Trustworthy inferred
● renzel ★★ Very Little Trustworthy
○ goekhan ★ Untrustworthy

Building a community to detect fake images in social networks.

Fig. 4 The fake media detection portal

With increasing community expertise either in form of points or achievements, players can reach four different expertise levels (i.e. novice, intermediate, advanced, and expert). Initially, a newcomer to the system has no points and achievements and is thus assigned to the lowest level *novice*. In the further process the player is encouraged to earn more points and achievements to reach the next level. Agents are shown their current levels and the number of points they need to reach the next level or to win achievements. Presenting the current expertise levels additionally motivates agents to continue rating media. Moreover, they can see other agents currently connected to the system and their status, which offers some kind of transparency; every one sees the same thing [16]. Furthermore, with increasing expertise, players can unlock special rights to upload and share more media to have them evaluated by other community members, which makes them more influential in the fake media detection community.

Explanations of all collectable achievements and the conditions for earning them must be visible to the player at all times. Normal achievements are awarded to agents reaching a certain number of authenticity ratings over a given period of time, upload a certain number of media for different communities or media of particular category (political, science, animals, etc.) and other normal achievements for rating a certain number of media for different communities. For certain special actions motivational symbolic achievements can be attached to the player's profile and reflected showing the corresponding media. Examples for such actions and achievements are “Be the first one to rate this image” (Initiator), “Be among the first ten agents to rate this image” (Pioneer), etc. Special achievements are then granted to players for a certain number of trustable media authenticity ratings, e.g. for having submitted 100 trustable authenticity ratings or for uploading 50 pictures in a special category.

Each increase in community activity or expertise points or achievements further increases the player's chance to be ranked higher in top-N expert lists for particular communities or in general, where community expertise is weighted much higher than mere community activity. Top-N ranking lists can show high scores achieved on a daily, weekly, monthly or all-time basis.

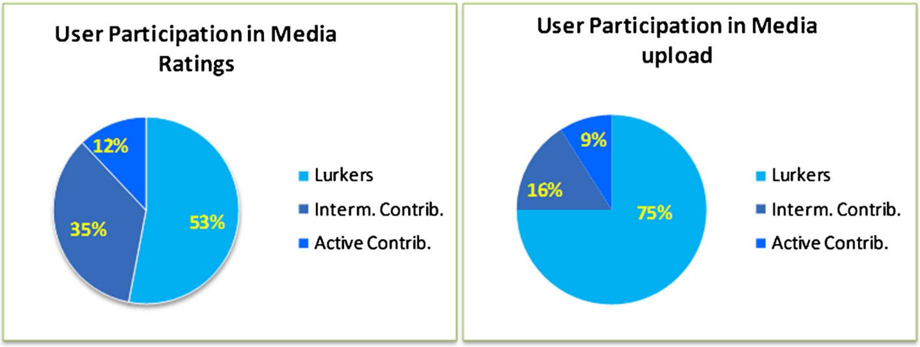
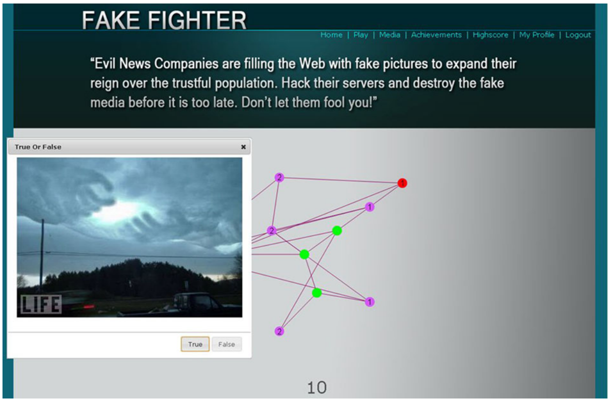


Fig. 5 Game evaluation


FAKE FIGHTER
Home | Play | Media | Achievements | Highscore | My Profile | Logout
True Or False
LIFE
True
False
10

“Evil News Companies are filling the Web with fake pictures to expand their reign over the trustful population. Hack their servers and destroy the fake media before it is too late. Don’t let them fool you!”

Fig. 6 The fake media detection mini-game “fake fighter”

3.2 Game evaluation

We found that using a game approach can be an effective way to collect data. We have collected a set of suspected media uploaded to our system in a short period of time, and a large number of media authenticity ratings were given by players.

According to the players’ comments, the game seems to be fun and enjoyable. In a 6 day period, close to 45 users joined the community, and contributed about 120 suspect images. Via the rating game, over 1,300 individual user ratings for these suspect images were collected. As an analogy to the 90-9-1 rule, we have created charts depicting user contributions in our community portal.

Figure 5 demonstrates the fraction of members belonging to each of the three categories: heavy contributors, intermittent contributors and lurkers. 9% of the members were responsible for the contribution of a majority of the suspect media, 16% uploaded a few whereas 75% did not upload any media. Concerning the media ratings, the percentage of very active members in terms of providing ratings and playing the game is 12%, of active members is 35% and of inactive members (few or no ratings) is 53%. We are intending to evaluate the game again after a longer period of time.

The game we have presented is the first prototype of a mini-games series. In the meantime, a set of further fake media detection mini-games was developed introducing new concepts such as hacking through a network of malicious media agencies to also make players aware of media distribution (mal)practices (cf. Fig. 6).

4 Trust management in collaborative fake media detection

The power of community information systems comes from their ability to capture real-world phenomena such as collaboration, competition and partnerships. In general, trust is an essential issue in all such systems involving interactions between people. A lack of information about community member reputation such as their credibility, reliability, and finally their trustworthiness can lead to distrust within and across communities. Consequently, the concept of trust must not be neglected in our approach of collaborative fake media detection. When it comes to users rating the authenticity of media and other users having to rely on such ratings, then a Multimedia Quality Profile derived from community authenticity ratings must also contain reputation information about the involved raters. Without this information the system becomes attackable easily. For example if the average of all media authenticity ratings is used as the only proxy measure for media authenticity, then it is easy to influence in a wrong direction by programmatically creating a huge amount of user accounts and letting all of these accounts rate maliciously [15, 82, 83].

Thus, especially in our proposed approach of collaborative fake media detection games, support for trust management and awareness must be an integral part of the system. It should be noted that we deal with trust not from the classical security point of view, but from a behavioral and social perspective. Approval of the agents' trustworthiness is not established on static certificates again issued by authorities that must be trusted. It rather relies on outcomes of agents' previous interactions in terms of reliability.

In this section, we discuss our approach to trust management and awareness in collaborative fake media detection systems in more detail. First, we discuss key characteristics and challenges in trust management we found in literature. Then, we present our refined model and implementation of trust management and awareness. Furthermore, we present an evaluation of trust inference in collaborative fake media detection system and discuss evaluation results.

Characteristics of trust Trust has been studied in several disciplines, such as sociology, psychology and economics, and definitions greatly vary across disciplines. Although until now there is no consensus on a complete definition of trust, there is wide consensus that trust is essential regarding interactions among people. First of all, trust is a key to positive interpersonal relationships in various settings [19, 43]. As a consequence, trust is key for the initiation of cooperative endeavors [5, 14, 20]. Trust becomes even more central and critical in cases of uncertainty due to organizational crisis [54].

Computationally, trust between entities is modeled as an overlay network of the original social graph, where edges are augmented with specific *trust values* of one entity towards the other. *Trust networks* are thus (partial) social networks in which the entities augment the edges by expressing their trust towards other entities. In most works on computational trust, these values are normalized to the range of 0 (no trust) to 1 (full trust). Victor et al. [59] state that there exist two important problems influencing trust opinions, which are challenges for trust networks. First, in large networks, it is likely that many agents do not know each other, which causes an abundance of ignorance. Second, there often exists a lack of a central authority, different agents might provide different and even contradictory information. These two issues may affect trust estimations.

No matter what the particular definitions of trust are, all of them refer to a basic set of key characteristics [25, 27, 33, 52] we considered in modeling and working with trust networks.

Transitivity Trust can be transitive. Let agents A and B trust each other well and B has a friend C. Then it is very likely that A may trust C to a certain extent even though they have never met. However, it is also true and reasonable for A not to trust C referring to that trust is not perfectly transitive in the mathematical sense, especially with growing path length between individuals in the social graph. Intuitively, trusting a friend's friend is less reliable than trusting a direct friend or even myself. Therefore, it is reasonable to say that trust decreases with increasing path length.

Asymmetry Trust relationships are asymmetric. While A may trust B to 100%, B may not necessarily exhibit the same trust towards A.

Personalization Trust is relative and dependent on personal subjective opinion, often based on previous experiences. Two agents may thus have very different opinions about the trustworthiness of the same other agent. Furthermore, an agent might give very little consideration to the opinions of agents he does not trust much, while the opinions of agents he strongly trusts are given more consideration.

Context-dependency Trust is context dependent. An agent can possibly trust a colleague sufficiently in a working setting, while he does not trust him very much in a private setting.

Dynamicity Trust is dynamic and can change over time with additional experiences. While an agent might have been trustworthy for a long time, a series of malicious actions can harm his trustworthiness for the future and even cause distrust (cf. Section 4.1).

From the network perspective, trust metrics are subdivided into *global* and *local* metrics [88]. Global trust metrics take into account all network members and trust relationships connecting them. In global trust computation, each agent's trustworthiness is computed from the perspective of the whole network. Each agent is thus assigned a single trust value. Global trust ranks are assigned to an individual based upon complete trust graph information. Such metrics are introduced in [25, 41]. Local trust metrics take into account personal bias as personalized trust [55, 56]. Such metrics take the agent for whom to compute trust as an additional input parameter and are able to operate on partial trust graph information. Trust inference in local trust computation is done from the perspective of another agent. Therefore, each agent can have more than one trust value in the network.

A lot of trust modeling approaches exist, but completely ignore the concept of distrust (e.g. [42, 44, 55, 85]). The emergence of interest in modeling the notion of distrust has been highlighted in literature [18, 23, 75], but there still exist few models considering both trust and distrust. Some approaches consider trust and distrust as opposite ends of the same continuous scale "negative mirror-image of trust" (cf. [1, 24]). On the other hand, there are opinions arguing that distrust cannot be seen as a lack of trust [12, 21, 22, 50]. Gans et al. [22] state that *"Distrust is regarded as just the other side of the coin, that is, there is generally a symmetric scale with complete trust on one end and absolute distrust on the other"*. Moreover, the authors stress that distrust is a reducible phenomenon that cannot be offset against any other social mechanisms and that distrust is relevant to the formation of social networks. We share Gans's notion of trust accompanied by the additional concepts of *confidence* and *distrust* in his TCD model [21].

In addition to trust, users experience a certain level of confidence in systems managing trust, i.e. its technical realization, transparent rules of participation, operating institutions and overall population. The distinction between trust and confidence plays an important role for the regulation and control of social networks. Networks need to develop binding rules facilitating trust-based interactions between members, e.g. for protecting confidentiality of information, for sanctioning breaches of trust, etc. According to Gans, the implementation of such rules is essential for the long-term success of social networks. Although confidence and trust may facilitate increased cooperation among network members, this cooperation has its costs, since network members still have to be watchful towards their trusted peers. Such watchfulness usually results in monitoring and reflecting the trust in individual members and the confidence in the system as a whole.

Additionally, the concept of distrust causes watchfulness and monitoring. Gans defines distrust as expectation of opportunistic behaviour by partners. In the process of cooperation, individuals might observe irregularities in the behaviour of their peers, leading to an ascent of distrust towards them. Once reaching individual thresholds, individuals start investigating on irregularities as evidence of their peers not being trustful. With a low distrust level, each irregularity might be seen as a single unplanned event, while with high distrust level, irregularities are seen as a series of planned negative events. On the first glance, distrust might appear as a negative mental state of one individual towards its peers, but the contrary is true. Gans proved that distrust vitalizes the network culture in terms of introducing a constant feeling of insecurity, thus resulting in more flexible and attentive behaviour of its participants.

In the remainder of this section, we discuss how we applied the concepts of trust, confidence and distrust in fake media detection, in particular during the collection of media authenticity ratings being the main input for our Multimedia Quality Profile.

4.1 Trust, confidence, distrust & media authenticity ratings

In [63] we presented a simplified model of today's media distribution based on a Publish-Subscribe model. A so-called *mediator* m maintains a list of sources S_m proposing media for publication. In order to provide better decision support for a mediator for or against the actual publication of an information item i and in turn to prevent reputation loss of m due to publishing media item i as true although it was a fake or vice versa, we introduced the notion of trust $t(m, s)$ of m towards any of his sources $s \in S_m$ into the computation of an aggregated authenticity rating $a(m, i)$ of m towards i over all submitted ratings $r(i, s)$.

$$a(m, i) = \frac{\sum_{j=1}^{|S_m|} t(m, s_j) \cdot r(i, s_j)}{\sum_{j=1}^{|S_m|} t(m, s_j)} \tag{1}$$

Furthermore, we presented a set of Web services enabling users to assign trust values to their contacts to provide authenticity ratings for media. We also described an algorithm that adapted trust values based on agreement or disagreement between a respective mediator and its sources [63]. On the event of mediator m taking either one of the actions of publishing a medium i as fake ($p_{\text{fake}}(m, i)$) or real ($p_{\text{fake}}(m, i)$), the trust values of m towards all $s \in S_m$ was automatically raised, if m and s were agreeing resp. lowered, if m and s were disagreeing.

However, in [63] we made a couple of oversimplified assumptions that needed refinement:

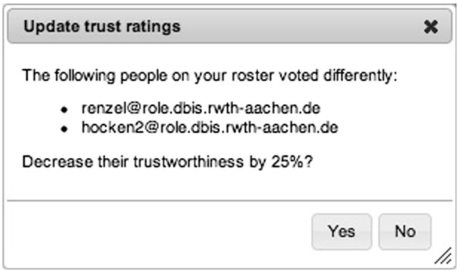
Trust inference We assumed that all trust values $t(m, s)$ were available. However, in a realistic scenario a mediator does in many cases not have explicit information about the trustworthiness of a yet unknown source in the moment of establishing a relationship or having to quickly decide about publication. In such cases m can only assume standard initial trust values towards s, e.g. being optimistic ($t(m,s) = 1$), pessimistic ($t(m, s) = 0$) or indefinite (e.g. $t(m, s) = 0.5$ or not defined). As a consequence, the results for $a(m, i)$ are comparably imprecise. It is thus desirable to derive improved a-priori information on trust values for yet unknown contacts to improve the precision of authenticity ratings. In our work we addressed this problem by introducing a modified version of Golbeck's TidalTrust trust inference algorithm [25]. Our implementation is described in more detail later in this section.

Trust bootstrapping Another challenge we faced is *trust bootstrapping*. The challenges are how to predict trust between agents in case no interactions have been captured so far and how to integrate newcomers which have no relations to any other members in the system in an existing trust network major component. In collaborative systems, trust is highly connected to the users' capability, activity and interest similarity. If a user's assigned ratings are similar to well-known trusted users, they may enjoy initial trust to some extent. According to [73], there is strong evidence that users who are similar-minded tend to trust each other more than any random ones. For example movie recommendations of people with same interests are usually more trustworthy than the opinions of people with a different taste. Ziegler and Golbeck [87] have proved a correlation between user similarity and trust. Thus, trust can be predicted by similarity to a certain extent. However, trust prediction by rating similarity is not yet implemented in our system.

Dynamicity of trust & authenticity ratings Whenever an aggregated authenticity rating is computed, it only reflects the current situation, i.e. the current levels of trust of m towards its sources S_m and the current ratings of each source. However, trust is highly dynamic and is changing over time, often in conjunction with significant events. Furthermore, sources might change their opinion about the authenticity of a medium at a later stage. Thus, the temporal dimension must be explicitly adopted into our equation. Consequently, our Media Quality Profile must be considered dynamic as well. Practically, each trust rating towards an entity should be accompanied by the timepoint of rating. With this combination, trust progression becomes traceable (cf. Section 4.2).

Automatic vs. manual trust updates Our trust update algorithm suggested to adapt trust values automatically. However, it was reported by users acting as mediators, that they prefer to stay in control of adapting trust values. Our algorithm is still used to make the mediator aware of potentially untrustable sources and to propose, but not automatically enforce a negative trust update (cf. Fig. 7).

Trust, confidence & distrust We assumed that trust is the only concept playing a role in relationships between mediators and their sources. However, we agree with Gans' TCD model and thus extended the existing model by the two additional


Update trust ratings

The following people on your roster voted differently:

- renzel@role.dbis.rwth-aachen.de
- hocken2@role.dbis.rwth-aachen.de

Decrease their trustworthiness by 25%?

Fig. 7 Making mediators aware of potentially untrustable sources

concepts of confidence and distrust. First, trust among community agents must be represented in authenticity ratings as precise as possible. Second, confidence in the system managing authenticity ratings and trust and producing Media Quality Profiles in the end must be given in terms of security and privacy protection on the one hand and reliability and validity of the underlying models and algorithms on the other hand. Finally, distrust must be supported by means of transparency regarding the rules used for trust and rating inference as well as awareness of other agents' activities. In our notion, the established level of confidence controls the level of system usage and the number of involved users, while trust and distrust control the density of the emerging social network of mediators and sources, in extreme cases possibly leading to percolation (cf. [6, 9, 13]). Data sources for powering the inclusion of confidence and distrust proxy measures into our model already exist, e.g. in form of the MobSOS/X monitoring data model [62, 70]. In conjunction with data mining and visualization tools applied on such a dataset it is theoretically possible to create awareness for the confidence of users in the whole system or its individual services as well as to support watchfulness in terms of monitoring individual users' activities with respect to the distrust factor. However, concrete proxy measures and computation rules thereof are still subject to further research.

Taking the above considerations into account, we refined our initial model for aggregated media authenticity ratings (2).

$$a(m, i, t) = \zeta(m, t) \frac{\sum_{s \in S_m} \tau(m, s, t) \cdot \delta(m, s, t) \cdot r(i, s, t)}{\sum_{s \in S_m} \tau(m, s, t) \cdot \delta(m, s, t)} \tag{2}$$

All functions referring to dynamic concepts such as trust, confidence, and distrust, as well as authenticity ratings were augmented with an additional parameter t modeling the timepoint of sampling. $\zeta(m, t)$ models the confidence of a mediator m in the whole system at a timepoint t. $\tau(m, s, t)$ resp. $\delta(m, s, t)$ model the level of trust resp. distrust a mediator m maintains towards one of his sources $s \in S_m$ at a given timepoint t. Besides the mere computation rules, the system provides technical support for these concepts. In the following paragraphs, we describe our technical realization.

Support for trust management & inference In [63] we presented a model relating media authenticity ratings with trust values of community agents as well as an algorithm that adapted trust values based on the agents' activity history. What remained unclear were initial trust values towards unknown agents—only uninformed initial presets such as being pessimistic, optimistic or undecided were possible. We argue that with a trust inference mechanism, we can improve the precision of $a(m, i, t)$ by more informed trust ratings for yet unknown agents. Based on the properties characterizing social networks [4, 53, 80] and a literature study on trust inference [25, 34, 51, 66, 84], we implemented a slightly modified version of the TidalTrust algorithm by Golbeck [25] for social networks induced by XMPP roster lists [68] as a LAS [74] Web service. TidalTrust uses the intuitive notion of asking direct or indirect contacts about their opinion on the trustworthiness of the unknown source. Technically, the algorithm first searches for the shortest paths between two entities in the trust network graph in a breadth-first-search (BFS) manner and then selects a set of paths for which an aggregated trust value is computed. However, in contrast to Golbeck, our algorithm works on directed graphs. The other difference to Golbeck's algorithm is that we dropped the constraint of a longest path length threshold due to the complex network properties inherently found in such networks. Since the basic approach of TidalTrust is BFS, the worst-case time complexity is in $O(|V| + |E|)$. However, the special network properties and initial simulation results suggest a much better practical time complexity for healthy trust networks.

Practically, we enabled the augmentation of any XMPP entity by a trust value and used XMPP roster lists [68] to form a trust network. Thus, we can manage and infer trust values for literally any XMPP entity (individual persons, communities, automatic agents, servers, etc.). In its current state, each XMPP entity transfers its complete roster list to a Web service responsible for storing and infering trust values over a complete network of XMPP entities. Thus, it would be possible to compute global trust values of all users. However, we believe that trust information should rather be kept distributed, so that organizations or individuals remain in control and decide their own policies. In practice, such a distribution is only achievable with additional protocols. Following our approach of using the XMPP protocol, it is feasible to directly integrate trust information with roster lists maintained at the respective server installation of an organization or an individual.

Support for confidence & distrust We agree with Gans' TCD model and thus do not consider distrust as the exact opposite of trust in the sense that trust and distrust add to a zero sum. While the level of trust reflects the level of trustful actions in terms of willingness to commit valid authenticity ratings, the level of distrust encourages the engagement in monitoring the sources providing authenticity ratings.

Confidence in the system itself in terms of security and privacy protection is guaranteed in the XMPP core specifications [67] with features such as secure authentication via SASL and TLS channel encryption. Confidence in the underlying models and algorithms is provided by public documentation and can in future be augmented by visualizations of the algorithm's basic functioning principles (cf. [28, 29, 58]). Using the game-based approach described in Section 3, such visualizations can be added to the game UI. In order to provide awareness on other agent's activities, we employ MobSOS/X monitoring [62, 70] to derive and visualize aggregated activity histories. Again, visualizations can be shown in interfaces such as the game UI described in Section 3.

4.2 Evaluation of trust management

In a short evaluation study involving 14 subjects, we evaluated a prototype fake media detection system enabling users to assign trust ratings to other users as well as to upload and assign authenticity ratings to media (cf. Fig. 4). Subjects were divided in two groups. A "good group" was instructed to rate correctly to the best of their knowledge, while an "evil group" was instructed to rate maliciously wrong. Subjects were instructed to not expose their role during the whole evaluation session. Our prototype additionally implemented intent-based realtime communication between the indidual widgets establishing the fake media detection UI used on the side of participants. We also used trust values and thresholds for enabling users to control if intents are allowed to change the user's view or should be dropped. During the evaluation period in total 176 trust ratings have been submitted for 104 distinct relationships between XMPP users. Eighty-eight of the distinct relations were established between participants of the evaluation session. In the same time period, 6259 trust ratings have been inferred by our service for trust management. The high number of inferences can be explained with the number of intents that have been published during the use of the *Fake Media Detection Tool*. Since the trust service has been configured to infer ratings on the fly, each recipient of an intent requested an inferred trust rating from the service. Moreover, the trust service was in theory able to infer trust ratings between any participant and any XMPP user since the *diameter* of the resulting graph is finite. As a result of the finite diameter, paths exist between any two XMPP users. However, for the group of participants, only $224 = 14 \times 16$ out of $240 = 15 \times 16$ possible inferences (without self-loops) could be computed because one participant did not provide trust ratings for any of his contacts. In total, trust ratings for 380 distinct pairs of XMPP users have been inferred. The distribution

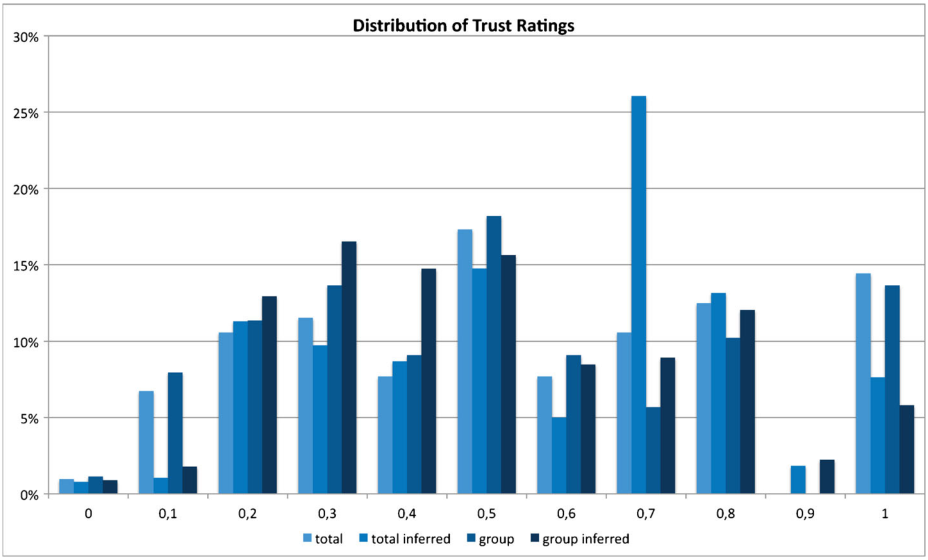


Fig. 8 Distribution of trust ratings

of trust ratings, subdivided into *assigned trust ratings in the total network*, *assigned trust ratings in the group of participants*, *inferred trust ratings in the total network*, and finally *inferred trust ratings in the group of participants* is presented in Fig. 8. Moreover, the progression of average trust ratings of the good and evil groups was analyzed. Note that trust ratings have been inferred from the viewpoint of the good group. Figure 9 depicts the progression of personal trust ratings for members of the good and the evil groups. The x-axis in both plots represents time in minutes. The y-axis represents the current average of trust ratings at a given timepoint. What is clearly visible is that the progression of the evil group reflects a quick decay of trust from the majority of other users. In other words, malicious users could be identified very quickly and precisely. The average trust ratings for each group depicted in Fig. 8 in the given order are 0.54, 0.57, 0.51, and 0.50. The average trust ratings

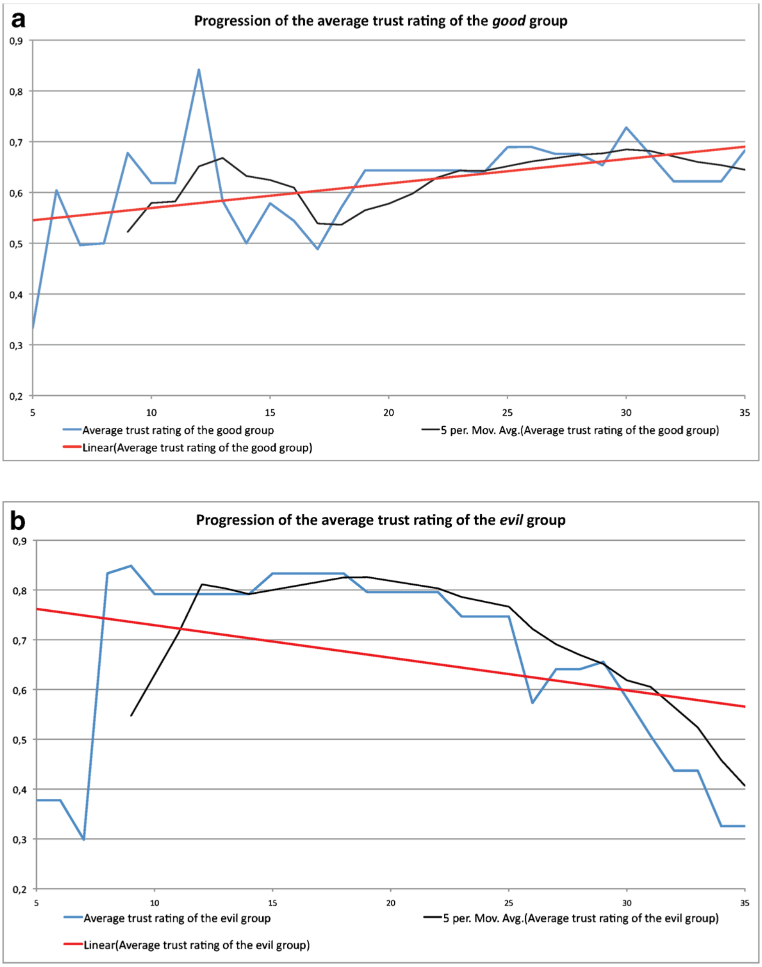


Fig. 9 Progressions of average trust ratings for a good and an evil group

thereby differ from the average trust rating that Golbeck has observed in the Film Trust network [25]. She monitored an average inferred trust rating of 0.7 throughout the network, which significantly differs from the null hypothesis of 0.5 and which is explainable with the increased trust that is usually put into acquaintances.

However, Golbeck's observations could be proven for certain time periods in the evil and good groups. Trust ratings inferred from members of the good group towards members of the evil group in the first 15 min of the session were very high (around 0.8). The reason is that all members of the evil group are usually well-trusted within the evaluator population. As a consequence, high trust ratings had been assigned to them by other participants who were not aware of their group membership in the beginning. During the evaluation session, the average rating of the evil group dropped to a value of 0.4. Their untrustworthy behavior was identified by subjects of the good group and then decreased as proposed by the trust service. Thus, the assumption that the average trust rating of the evil group decreases over time could be confirmed.

Golbeck's observations could again be proven for trust ratings that have been inferred at the end of the session between members of the good group. The trust rating is around 0.35 in the beginning, but increases to a value of 0.68 at the end, which complies with Golbeck's observations. The reason for the increase is the trustworthy behavior of the good group. Since members of the group mostly rated the authenticity of a particular image similarly, the assigned trust ratings between members of the good group increased, which in turn had an impact on inferred trust ratings computed by the trust service.

Golbeck's observation could not be proven for the total group of XMPP users nor for the group of participants of the session based on the trust ratings that were present after the session. The average trust rating of each group depicted in Fig. 8 is near the null hypothesis. A reason might be the separation of participants into good group and evil groups. Since the distribution of trust ratings in Fig. 8 has been determined after the session was finished, personal trust ratings for members of the evil group might not have been restored by all participants to the initial value.

5 Conclusions and future work

Faked multimedia are here to stay. With the enormous amount of multimedia materials produced and shared on the Web, the task of defining the authenticity of multimedia can only be solved by smart combinations of automatic methods and with the help of an overwhelmingly good-willing community of Web users which are in principle willing to identify fake multimedia since the reputation loss with the distribution of faked multimedia is threatening our confidence in the Web as a medium itself. A Media Quality Profile is proposed to be used for media classification. Such a profile is like a certificate, but its nature is highly dynamic. Even if quick decisions are often needed in media press agencies to judge if an image is faked or authentic, the Web community can do more research on a multimedia artifact and change the judgment. By doing so, experts and media are transferred in a network of trust relationships. This network can be evaluated and utilized in later decision making processes. However, certain practical issues need to be addressed by all platforms supporting the idea of community-based fake multimedia detection.

Typical cold-start problems in community information systems have to be addressed. Games are one way to motivate users to participate, and participation is the key in any Web-based community. Our game approach has its limitations and many other games for the identification of faked multimedia can be developed. From previous research and from an extensive literature review we have refined the idea of trust inference in networks of agents. We presented a Web service for inferring agents' trust and initial results of simulation runs with the service. Next steps include the integration of more media metadata in the quality profile, either embedded in MPEG-7 or other hybrid formats. Refinements of the game concept and metadata visualization are also on our agenda together with the security considerations for our proposed quality profile. More important is the location of experts in networks for identifying certain multimedia artifacts in special domains like war propaganda using variants of the HITS algorithm [38] and counter-measures against compromising the system which are known already from search engines in many cases. The games we have presented are the first prototype of a game series. We intend to extend and improve this scenario in several directions of our system.

Acknowledgements The research leading to these results has received funding from B-IT Research School and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 231396 (ROLE project).

References

1. Abdul-Rahman A, Hailes S (2000) Supporting trust in virtual communities. In: Proceedings of the 33rd Hawaii international conference on system sciences, vol 6. IEEE Computer Society, Washington, DC, USA, pp 1–7. http://portal.acm.org/citation.cfm?id=820262.820322
2. Ahn LV (2006) Games with a purpose. IEEE Comput Mag 39:92–94
3. Ahn LV, Dabbish L (2008) Designing games with a purpose. Commun ACM 51(8):58–67
4. Albert R, Jeong H, Barabási AL (1999) Diameter of the World-Wide Web. Nature 401(9): 130–131
5. Arrow KJ (1974) The limits of organization (Fels lectures on public policy analysis). W. W. Norton & Company
6. Bashan A, Parshani R, Havlin S (2011) Percolation in networks composed of connectivity and dependency links. Phys Rev E 83(5):1–8. doi:10.1103/PhysRevE.83.051127
7. Bober M, Brasnett P (2009) MPEG-7 visual signature tools. In: ICME'09: proceedings of the 2009 IEEE international conference on multimedia and expo. IEEE Press, Piscataway, NJ, USA, pp 1540–1543. ISBN 978-1-4244-4290-4
8. Brugion DA (1999) Photo fakery: the history and techniques of photographic deception and manipulation. Brasssyes Publishers, Dulles verginia
9. Callaway DS, Newman MEJ, Strogatz SH, Watts DJ (2000) Network robustness and fragility: percolation on random graphs. Phys Rev Lett 85(25):5468–5471. doi:10.1103/PhysRevLett.85.5468
10. Cheng R, Vassileva J (2005) User motivation and persuasion strategy for peer-to-peer communities. In: Hawaii international conference on system sciences, vol 7, p 193a. doi:10.1109/HICSS.2005.653
11. Clary EG, Snyder M, Ridge RD, Copeland J, Stukas AA, Haugen J, Miene P (1998) Understanding and assessing the motivations of volunteers: a functional approach. JPSP 74:1516–1530
12. Cofta P (2006) Distrust. In: Proceedings of the 8th international conference on electronic commerce, ICEC '06. ACM, New York, NY, USA, pp 250–258. doi:10.1145/1151454.1151498
13. Cohen R, Erez K, ben Avraham D, Havlin S (2001) Breakdown of the internet under intentional attack. Phys Rev Lett 86(16):3682–3685. doi:10.1103/PhysRevLett.86.3682
14. Deutsch M (1958) Trust and suspicion. JCR 2:265–279
15. Douceur JR (2002) The Sybil attack. In: Revised papers from the first international workshop on peer-to-peer systems, IPTPS '01. Springer, London, UK, pp 251–260

16. Erickson T (2003) Designing visualizations of social activity: six claims. In: CHI '03 extended abstracts on human factors in computing systems. ACM, New York, NY, USA, pp 846–847
17. Farid H (2009) A survey of image forgery detection. IEEE Sig Proc Mag 2(26):16–25. www.cs.dartmouth.edu/farid/publications/spm09.html
18. Finin T, Kagal L, Olmedilla D (eds) (2006) Proceedings of the WWW'06 workshop on models of trust for the Web (MTW'06), Edinburgh, Scotland, UK, 22 May 2006. CEUR Workshop Proceedings, vol 190. CEUR-WS.org
19. Fox A (1974) Beyond contract: work, power and trust relations. Faber
20. Gambetta D (1988) Can we trust trust?, chap. Trust: making and breaking cooperative relations. Blackwell, pp 213–237
21. Gans G (2008) An agent-based modeling-and simulation-methodology for strategic interorganizational networks. PhD thesis, RWTH Aachen University
22. Gans G, Jarke M, Kethers S, Lakemeyer G, Ellrich L, Funken C, Meister M (2001) Towards (Dis)Trust-based simulations of agent networks. In: Proceedings of the 4th workshop on deception, fraud, and trust in agent societies, Montreal, pp 49–60. http://www-i5.informatik.rwth-aachen.de/~gans/tropos/dokumente/Agent01.pdf
23. Golbeck J (2009) Computing with social trust (human-computer interaction series), 1st edn. Springer
24. Golbeck J, Parsia B, Hendler J (2003) Trust networks on the semantic web. In: Proceedings of seventh international workshop on cooperative intelligent agents (CIA-03), pp 238–249. http://www.mindswap.org/papers/Trust.pdf
25. Golbeck JA (2005) Computing and applying trust in web-based social networks. PhD thesis, University of Maryland, College Park
26. Golle P, Leyton-Brown K, Mironov I (2001) Incentives for sharing in peer-to-peer networks. In: Proceedings of the 3rd ACM conference on electronic commerce, EC '01. ACM, New York, NY, USA, pp 264–267
27. Grandison T, Sloman M (2000) A survey of trust in internet applications. IEEE Communications Surveys and Tutorials 3(4):2–16. http://pubs.doc.ic.ac.uk/TrustSurvey/
28. Gretarsson B, O'Donovan J, Bostandjiev S, Hall C, Höllerer T (2010) SmallWorlds: visualizing social recommendations. Computer Graphics Forum 29(3):833–842
29. Herlocker JL, Konstan JA, Riedl J (2000) Explaining collaborative filtering recommendations. In: Proceedings of CSCW00: ACM conference on computer supported cooperative work, pp 241–250
30. Hinduja S, Patchin JW (2008) Cyberbullying: an exploratory analysis of factors related to offending and victimization. Deviant Behav 29(2):129–156
31. Iwata M, Hori T, Shiozaki A, Ogihara A (2010) Digital watermarking method for tamper detection and recovery of JPEG images. In: International symposium on information theory and its applications (ISITA), pp 309–314
32. ISO/IEC/15938-3/Amd.3 (2009) Image signature tools
33. Jøsang A, Ismail R, Boyd C (2007) A survey of trust and reputation systems for online service provision. Decis Support Syst 43:618–644. doi:10.1016/j.dss.2005.05.019
34. Kamvar SD, Schlosser MT, Garcia-Molina H (2003) The eigentrust algorithm for reputation management in P2P networks. In: Proceedings of the 12th international conference on world wide web, WWW '03. ACM, New York, NY, USA, pp 640–651. doi:10.1145/775152.775242
35. Keith S, Martin ME (2005) Cyber-bullying: creating a culture of respect in a cyber world. RCY Journal 13(4):224–228
36. Kim AJ (2000) Community building on the web: secret strategies for successful online communities. Peachpit Press
37. Klamma R, Spaniol M, Renzel D (2007) Community-aware semantic multimedia tagging—from folksonomies to commsonomies. In: Tochtermann K, Maurer H, Kappe F, Scharl A (eds) Proceedings of I-Media '07. International conference on new media technology and semantic systems, Graz, Austria, 5–7 Sept 2007. J.UCS (Journal of Universal Computer Science) Proceedings, pp 163–171
38. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46:604–632. doi:10.1145/324133.324140
39. Kuznetsov S (2006) Motivations of contributors to Wikipedia. ACM SIGCAS Comput Soc 36(2):1–7

40. Lee S, Shamma DA, Gooch B (2006) Detecting false captioning using common-sense reasoning. Digital Investigation 3(Supplement-1):65–70
41. Levien R (2003) Attack resistant trust metrics. PhD thesis, UC Berkeley, Berkeley, CA, USA
42. Levien R (2009) Attack-resistant trust metrics. In: Golbeck J (ed) Computing with social trust, human-computer interaction series. Springer, London Limited, pp 121–132
43. Lewis JD, Weigert AJ (1985) Trust as a social reality. Soci Force 63(4):967–985
44. Li XY, Gui XL (2009) A comprehensive and adaptive trust model for large-scale P2P networks. J Comput Sci Technol 24:868–882. doi:10.1007/s11390-009-9278-4
45. Lux M (2009) Caliph & Emir: MPEG-7 photo annotation and retrieval. In: ACM multimedia, pp 925–926
46. Lux M, Becker J, Krottmaier H (2003) Caliph & Emir: semantic annotation and retrieval in personal digital photo libraries. In: Proceedings of CAiSE '03 forum at 15th conference on advanced information systems engineering, pp 85–89
47. Manjunath BS (2002) Introduction to MPEG-7, multimedia content description interface. Wiley
48. Manjunath BS, Ohm JR, Vasudevan VV, Yamada A (2001) Color and texture descriptors. IEEE Trans Circuits Systems Video Technol 11(6):703–715. doi:10.1109/76.927424
49. Manovich L (2001) The language of new media. The MIT Press
50. Marsh S, Briggs P (2009) Examining trust, forgiveness and regret as computational concepts. In: Golbeck J (ed) Computing with social trust, human-computer interaction series. Springer, London Limited, p 1
51. Massa P, Bhattacharjee B (2004) Using trust in recommender systems: an experimental analysis. In: Proceedings of iTrust2004 international conference, pp 221–235
52. Mcknight DH, Chervany NL (1996) The meanings of trust. Tech. rep., University of Minnesota, USA
53. Milgram S (1967) The small world problem. Psychol Today 1(1):61–67
54. Mishra AK (1996) Trust in organizations: frontiers of theory and research, chap. Organizational responses to crisis: the centrality of trust. Sage Publications, Inc, pp 261–287
55. Mui L (2002) Computational models of trust and reputation: agents, evolutionary games, and social networks. PhD thesis, Massachusetts Institute of Technology
56. Mui L, Mojdeh Mohtashemi AH (2002) A computational model of trust and reputation. In: Proceedings of the 35th Hawaii international conference on system science (HICSS). IEEE Computer Society, Washington, DC, USA, pp 2431–439. http://portal.acm.org/citation.cfm?id=820745.821158
57. Nielsen J (2006) Participation inequality: encouraging more users to contribute. http://www.useit.com/alertbox/participation_inequality.html. Last access: June 2011
58. O'Donovan J, Smyth B, Gretarsson B, Bostandjiev S, Höllerer T (2008) PeerChooser: visual interactive recommendation. In: Proceedings of CHI08: twenty-sixth annual SIGCHI conference on human factors in computing systems. ACM, pp 1085–1088
59. Victor P, Cornelis C, Cock MD (2011) Trust networks for recommender systems. Atlantis Press
60. Rafelsberger W, Scharl A (2009) Games with a purpose for social networking platforms. In: Proceedings of the 20th ACM conference on hypertext and hypermedia, HT '09. ACM, New York, NY, USA, pp 193–198
61. Rashed KAN, Klamma R (2010) Towards detecting faked images. In: Carreras A, Delgado J, Maronas X, Rodriguez V (eds) Proceedings of the 11th international workshop on interoperable social multimedia applications (WISMA10), Barcelona, Spain, 19–20 May 2010. CEUR Workshop Proceedings, vol 583
62. Renzel D, Klamma R, Spaniol M (2008) MobSOS–a testbed for mobile multimedia community services. In: Proceedings of the 2008 ninth international workshop on image analysis for multimedia interactive services. IEEE Computer Society, Washington, DC, USA, pp 139–142
63. Renzel D, Rashed KAN, Klamma R (2010) Collaborative fake media detection in a trust-aware real-time distribution network. In: Proceedings of the the 2nd workshop focusing on semantic multimedia database technologies SMDT2010, Saarbrucken, Germany. CEUR Workshop Proceedings, vol 680
64. Resnick P, Kuwabara K, Zeckhauser R, Friedman E (2000) Reputation systems. Commun ACM 43:45–48
65. Rey C, Dugelay JL (2002) A survey of watermarking algorithms for image authentication. EURASIP J Appl Signal Process 2002(1):613–621

66. Richardson M, Agrawal R, Domingos P (2003) Trust management for the semantic web. In: Fensel D, Sycara K, Mylopoulos J (eds) The Semantic Web—ISWC 2003. Lecture notes in computer science, vol 2870. Springer Berlin/Heidelberg, pp 351–368
67. Saint-Andre P (2004) RFC 3920—Extensible Messaging and Presence Protocol (XMPP): core. Tech. rep., Jabber Software Foundation. http://www.ietf.org/rfc/rfc3920.txt
68. Saint-Andre P (2004) RFC 3921 extensible messaging and presence protocol (XMPP): instant messaging and presence. Tech. rep., Jabber Software Foundation. http://www.ietf.org/rfc/rfc3921.txt
69. Schein A, Popescul A, Ungar L, Pennock D (2002) Methods and metrics for cold-start recommendations. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2002), pp 253–260
70. Schlebusch P (2009) XMPP based monitoring and analysis of mobile communities. Master's thesis, RWTH Aachen University
71. Shneiderman B (2004) Designing for fun: how can we design user interfaces to be more fun? Interactions 11(5):48–50. doi:10.1145/1015530.1015552
72. Shneiderman B, Plaisant C (2004) Designing the user interface: strategies for effective human-computer interaction, 4th edn. Bosten, Addison Wesley
73. Skopik F, Schall D, Dustdar S (2009) Start trusting strangers? Bootstrapping and prediction of trust. In: Proceedings of 10th international conference WISE, Poznan, Poland, 5–7 Oct 2009. Springer, pp 275–289
74. Spaniol M, Klamma R, Janßen H, Renzel D (2006) LAS: a lightweight application server for MPEG-7 services in community engines. In: Tochtermann K, Maurer H (eds) Proceedings of I-KNOW '06, 6th international conference on knowledge management, Graz, Austria. J. UCS (Journal of Universal Computer Science) Proceedings. Springer, pp 592–599
75. Stølen K, Winsborough WH, Martinelli F, Massacci F (eds) (2006) Trust management. In: Proceedings of the 4th international conference, iTrust 2006, Pisa, Italy, 16–19 May 2006. Lecture Notes in Computer Science, vol 3986. Springer Berlin Heidelberg
76. Susi T, Johannesson M, Backlund P (2007) Serious games—an overview. Tech. rep., School of Humanities and Informatics University of Skvde, Sweden
77. Technical Standardization Committee on AV & IT Storage Systems and Equipmen (2002) Exchangeable image file format for digital still cameras: Exif Version 2.2. Tech. rep., Standard of Japan Electronics and Information Technology Industries Association
78. Turner R (2007) Your online community beta program: Metcalfe's law and the cold start. http://www.websocialarchitecture.com/community/2007/09/. Last access: June 2011
79. Wang RY, Reddy MP, Kon HB (1995) Toward quality data: an attribute-based approach. Decision Support Syst 13:349–372. http://web.mit.edu/tdqm/www/tdqmpub/Toward%20Quality%20Data.pdf
80. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393(6684):440–442
81. Webster A, Vassileva J (2006) Visualizing personal relations in online communities. In: Lecture notes in computer science adaptive hypermedia and adaptive webbased systems, pp 223–233. doi:10.1007/11768012_24
82. Wu B, Davison BD (2005) Identifying link farm spam pages. In: Special interest tracks and posters of the 14th international conference on world wide web. ACM, New York, NY, USA, pp 820–829
83. Yang Y, Sun YL, Kay S, Yang Q (2009) Defending online reputation systems against collaborative unfair raters through signal modeling and trust. In: Proceedings of the 2009 ACM symposium on applied computing, SAC '09. ACM, New York, NY, USA, pp 1308–1315
84. Yaniv I, Kleinberger E (2000) Advice taking in decision making: egocentric discounting and reputation formation. OBHDP Journal 83(2):260– 281
85. Zaihrayeu I, da Silva PP, McGuinness DL (2005) IWTrust: improving user trust in answers from the web. In: Proceedings of the 3rd international conference on trust management (iTrust2005), pp 384–392
86. Zajonc RB (1965) Social facilitation. Science 149:269–274
87. Ziegler CN, Golbeck J (2007) Investigating interactions of trust and interest similarity. Deci Supp Syst 43:460–475. doi:10.1016/j.dss.2006.11.003
88. Ziegler CN, Lausen G (2005) Propagation models for trust and distrust in social networks. Inf Syst Front 7:337–358. doi:10.1007/s10796-005-4807-3
89. Zyda M (2005) From visual simulation to virtual reality to games. IEEE Comput Mag 38:25–32

Khaled Ahmed Nagi Rashed received his bachelor and master degrees in computer engineering and technologies in 1996, and 1998, respectively from the faculty of informatics and robotics, department of computer technology and information security, USATU University, Russia. Currently, he is a PhD candidate and a research assistant at department of Computer Science, information systems chair, RWTH Aachen University, Germany. His research interests are multimedia information systems, multimedia metadata, community information systems, trust and reputation systems for Online collaboration and web mining.

Dominik Renzel received his diploma degree in computer science in February 2009 from RWTH Aachen University. He is working as a research assistant and PhD student at the Chair of Computer Science 5 at RWTH Aachen University since March 2009. His research interests include Technology Enhanced Learning, Personal Learning Environments (PLE), Evaluation and Analysis of Mobile Multimedia Community Information Systems, Multimedia Metadata Management & Standards, and Protocols for Real-time Online Communication & Collaboration. Dominik is reviewer for the Journal Multimedia Tools & Applications (MTAP).

Ralf Klamma has diploma, doctoral and habilitation degrees in computer science from RWTH Aachen University. He leads the research group “community information systems” at the information systems chair, RWTH Aachen University. He is coordinating and working in four major EU projects for Technology Enhanced Learning (ROLE, TELMAP, GALA and TELLNET), He is member of the research excellence cluster “Ultra High Speed Mobile Information and Communication” (UMIC) specialized in mobile multimedia. Ralf organized doctoral summer schools and conferences in Multimedia Technology Enhanced Learning, and Social Network Analysis. He is on the editorial board of IEEE Transactions on Technology Enhanced Learning and Social Networka Analysis and Mining (SNAM). His research interests are community information systems, multimedia metadata, social network analysis and technology enhanced learning.

Matthias Jarke is Professor of Information Systems at RWTH Aachen University and Executive Director of the Fraunhofer FIT Institute for Applied Information Technology. He is founder director of the Bonn-Aachen International Graduate Center for Information Technology (B-IT) which is supported by RWTH Aachen, the University of Bonn, and the Fraunhofer-Gesellschaft. Jarke holds master degrees in business administration and computer science and a doctorate in business informatics, both from the University of Hamburg, Germany. Prior to joining Aachen, he held faculty positions at New York UniversityŠs Stern School of Business and at the University of Passau. His research area is information systems support for cooperative activities in business, engineering, and culture. He has been coordinator of three European research projects in these areas. He is currently deputy coordinator of the German national excellence cluster in mobile communications in Aachen. Jarke has published about 25 books and over 250 refereed papers. Jarke was a Chief Editor of the journal Information Systems from 1993–2003, and has served as program chair of major international conferences such as VLDB, EDBT, CoopIS, and CAiSE. He is elected senior reviewer for software engineering for the German national science foundation DFG. From 2004–2007, he was president of the German Informatics society, GI.